Speech processing for multimodal and adaptive systems

HABILITATIONSSCHRIFT

ausgeführt zur Erlangung der Venia Docendi
für das wissenschaftliche Fach

"Speech Communication"

eingereicht an der
Technischen Universität Graz
Fakultät für Elektrotechnik und Informationstechnik

von

Mag.phil. Dipl.-Ing. Dr.techn. **Michael Pucher**
Institut für Schallforschung
Österreichische Akademie der Wissenschaften

Wien, Juni 2016

Kurzfassung

Diese Habilitationsschrift basiert auf Arbeiten im Bereich der *Sprachverarbeitung für multimodale und adaptive Systeme* die während der letzten 5 Jahre von mir und meiner Gruppe am *Forschungszentrum Telekommunication Wien* (FTW) durchgeführt wurden. Diese Arbeit konzentriert sich auf die drei Themen *textbasierte Sprachsynthese*, *audio-visuelle textbasierte Sprachsynthese* und der *Täuschung von Systemen zur Verifikation von SprecherInnen*. In diesen Gebieten zeigen wir wie adaptive und multimodale Ansätze Modellierungsalgorithmen und Methoden verbessern können. In der *textbasierten Sprachsynthese* zeigen wir wie ein akustisches SprecherInnenmodell an soziale und regionale Varietäten und Lautdauer adaptiert werden kann, und wie wir eine überwachte und unüberwachte Interpolation von Varietäten verwenden können um Zwischenvarietäten zu generieren. Diese Interpolation ermöglicht die Entwicklung von personalisierten Sprachdialogsystemen. Außerdem zeigen wir welchen Einfluß die Bekanntheit mit einer SprecherIn auf die Wahrnehmung von synthetischer Sprache hat. In der *audio-visuellen textbasierten Sprachsynthese* zeigen wir wie ein gemeinsamer audio-visueller Modellierungsansatz, bei dem akustische und visuelle Daten gemeinsam im Trainingsprozess verwendet werden, die audio-visuelle Modellierung verbessern kann. Mit einer derartigen Adaption können wir die audio-visuellen Trainingsdaten effizient verwenden, welche aus synchronen 3D Markersequenzen und akustischen Sprachaufnahmen bestehen. Wir zeigen auch wie ein visuelles Modell zur Kontrolle eines akustischen Modells verwendet werden kann und wie ein adaptiertes visuelles Modell die Synthese verbessern kann. In der *Täuschung von Systemen zur Verifikation von SprecherInnen* zeigen wir wie ein adaptives Sprachsynthesesystem ein Verifikationssystem täuschen kann, und wie dies durch ein wiederum adaptiertes Verifikationssystem vermieden werden kann. Seit der Publikation unserer Arbeit zu diesem Thema haben sich mehrere Forschungsgruppen mit dem Problem beschäftigt. Der Bereich der *Sprachverarbeitung für multimodale und adaptive Systeme* ist Teil des großen Forschungsbereichs der *Sprachkommunikation*. In diesem umfassenden Forschungsbereich hat es in den letzten Jahren einen enormen Fortschritt gegeben. Wir werden einen Überblick über dieses Feld geben das von der Sprachwahrnehmung, der Sprachsynthese und Erkennung bis zum Sprachverstehen reicht, und die verschiedenen Modalitäten in denen Sprache repräsentiert werden kann umfasst.

Abstract

This habilitation thesis is based on work in *Speech Processing for Multimodal and Adaptive Systems* that was conducted during the last 5 years by me and my group at *Telecommunications Research Center Vienna* (FTW). Our work is concentrated on the three topics *Text-to-Speech Synthesis* (TTS), *Audio-Visual Text-to-Speech Synthesis* (AVTTS), and *Speaker Verification Spoofing* (SVS) where we show how adaptive and multimodal approaches can improve state-of-the-art modeling algorithms and methods. In *Text-to-Speech Synthesis* we show how to adapt an acoustic speaker model to social and regional varieties and speech sound durations, and how we can apply supervised and unsupervised variety interpolation to generate in-between varieties. This interpolation allows us to develop a more flexible and personalized interaction in a spoken language interface. We also evaluate the effect of speaker familiarity on synthetic speech perception. For *Audio-Visual Text-to-Speech Synthesis* we show how a joint audio-visual modeling approach, where acoustic and visual data is used jointly in the training process, can improve the audio-visual modeling and how a visual control model can be used to control the acoustic model. We also investigate a speaker adaptive approach that can make efficient use of audio-visual training data, which consists of synchronous 3D marker sequences of facial movement and acoustic speech recordings. For *Speaker Verification Spoofing* we showed how adaptive speech synthesis systems can spoof speaker verification systems, and how adopting an adaptive speaker verification system can block this. Since the publication of our work, several research groups have started to work on the topic of *Speaker Verification Spoofing*. The topic of *Speech Processing for Multimodal and Adaptive Systems* can be embedded into the larger field of *Speech Communication* (SC). There has been a huge progress in this large field within the last years and we will sketch an overview of this field that ranges from speech perception, speech synthesis and recognition to spoken language understanding, including the different modalities in which speech can be represented.

Contents

1 Speech Communication **1**

1.1 Introduction . . . 1
1.1.1 Own contributions . . . 3
1.2 Speech Perception, Production and Acquisition . . . 3
1.2.1 Overview . . . 3
1.2.2 Own contributions . . . 4
Interaction speech production-speech perception . . . 4
1.3 Phonetics, Phonology, and Prosody . . . 4
1.3.1 Overview . . . 4
1.3.2 Own contributions . . . 5
Sociophonetics . . . 5
1.4 Analysis of Paralinguistics in Speech and Language . . . 5
1.4.1 Overview . . . 5
1.4.2 Own contributions . . . 6
Non-verbal communication . . . 6
1.5 Speaker and Language Identification . . . 6
1.5.1 Overview . . . 6
1.5.2 Own contributions . . . 7
Automatic Speaker Verification Spoofing and Countermeasures . . . 7
1.6 Analysis of Speech and Audio Signals . . . 7
1.6.1 Overview . . . 7
1.6.2 Own contributions . . . 7
Audio signal analysis and representation . . . 7
Pitch and harmonic analysis, Singing analysis . . . 8
1.7 Speech Coding and Enhancement . . . 8
1.7.1 Overview . . . 8
1.7.2 Own contributions . . . 9
1.8 Speech Synthesis and Spoken Language Generation . . . 9
1.8.1 Overview . . . 9
1.8.2 Own contributions . . . 10
Unit selection speech synthesis . . . 11
Statistical parametric speech synthesis . . . 11
Prosody modeling and generation . . . 12
Expression, emotion and personality generation . . . 13
Synthesis of singing voices . . . 13
Voice modification, conversion and morphing . . . 13

Avatars and talking faces . . . 13
Tools and data for speech synthesis . . . 14
Evaluation of speech synthesis . . . 15
1.9 Speech Recognition – Signal Processing, Acoustic Modeling, Robustness, Adaptation . . . 15
1.9.1 Overview . . . 15
1.9.2 Own contributions . . . 16
Acoustic modeling for conversational speech (dialog, interaction) . 16
1.10 Speech Recognition – Architecture, Search, and Linguistic Components . 16
1.10.1 Overview . . . 16
1.10.2 Own contributions . . . 17
Language modeling for conversational speech (dialog, interaction) . 17
1.11 Speech Recognition – Technologies and Systems for New Applications . . 17
1.11.1 Overview . . . 17
1.11.2 Own contributions . . . 18
Multimodal systems . . . 18
New paradigms (e.g. artic. models, silent speech interfaces, topic models) . . . 19
1.12 Spoken Language Processing – Dialog, Summarization, Understanding . . 19
1.12.1 Overview . . . 19
1.12.2 Own contributions . . . 20
Spoken dialog systems . . . 20
Multimodal human-machine interaction (conversat. agents, human-robot) . . . 20
Semantic analysis and classification . . . 20
Evaluation of speech and multimodal dialog systems . . . 21
1.13 Spoken Language Processing – Translation, Information Retrieval, Resources . . . 21
1.13.1 Overview . . . 21
1.13.2 Own contributions . . . 22

2 Speech processing for multimodal and adaptive systems 23
2.1 Introduction . . . 23
2.1.1 Text-to-Speech Synthesis . . . 24
Impact of our work on dialect speech synthesis . . . 24
Impact of our work on modeling of fast speech . . . 26
Impact of our work on perception of synthetic speech . . . 26
Artificial Intelligence (AI) and context modeling . . . 26
Realism in speech synthesis . . . 27
2.1.2 Audio-Visual Text-to-Speech Synthesis . . . 27
Impact of our work on audio-visual speech synthesis . . . 28
2.1.3 Speaker Verification Spoofing . . . 29
Impact of our work on speaker verification spoofing . . . 29

2.2 Text-to-Speech Synthesis . . . 31
2.2.1 Supervised variety interpolation . . . 31
2.2.2 Unsupervised variety interpolation . . . 34
2.2.3 Modeling of fast speech . . . 37
2.2.4 Perception of synthetic speech . . . 39
PAPER: Unsupervised and phonologically controlled interpolation of language varieties for speech synthesis . . . 43
PAPER: Intelligibility of time-compressed synthetic speech: compression method and speaking style . . . 61
PAPER: Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis . . . 75
PAPER: Influence of speaker familiarity on blind and visually impaired childrens perception of synthetic voices in audio games . . . 91
PAPER: Intelligibility analysis of fast synthesized speech . . . 97
PAPER: Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and sighted listeners . . . 103
2.3 Audio-Visual Text-to-Speech Synthesis . . . 107
2.3.1 Joint audio-visual modeling . . . 107
2.3.2 Adaptive audio-visual modeling . . . 109
2.3.3 Visual control of acoustic speech . . . 110
PAPER: Joint Audiovisual Hidden Semi-Markov Model-based Speech Synthesis . . . 113
PAPER: Speaker-adaptive visual speech synthesis in the HMM-framework 125
PAPER: Visual Control of Hidden-Semi-Markov-Model based Acoustic Speech Synthesis . . . 129
2.4 Speaker Verification Spoofing . . . 135
PAPER: Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech . . . 139

Principal references **151**

Secondary references **153**

Bibliography **161**

A Sample webpages **185**
A.1 Text-to-Speech Synthesis . . . 185
A.2 Audio-Visual Text-to-Speech Synthesis . . . 185

B List of principal references **187**
B.1 Text-to-Speech Synthesis . . . 187
B.1.1 Journals . . . 187
B.1.2 Conferences . . . 188
B.2 Audio-Visual Text-to-Speech Synthesis . . . 188

B.2.1 Journals 188
B.2.2 Conferences 188
B.3 Speaker Verification Spoofing 189
B.3.1 Journals 189

Curriculum Vitae **189**

List of Acronyms

3GPP *3rd Generation Partnership Project*

AI *Artificial Intelligence*

AM *Acoustic Model*

AMI *Augmented Multi-Party Interaction*

AMR *Adaptive Multi-Rate*

ASR *Automatic Speech Recognition*

AVTTS *Audio-Visual Text-to-Speech Synthesis*

CELP *Code Excited Linear Prediction*

CFG *Context Free Grammar*

CS *Computational Semantics*

CUI *Cognitive User Interfaces*

DFT *Discrete Fourier Transform*

DNN *Deep Neural Networks*

EER *Equal Error Rate*

EM *Expectation Maximization*

EMA *ElectroMagnetic Articulography*

EVS *Enhanced Voice Services*

FTW *Telecommunications Research Center Vienna*

G2P *Grapheme-To-Phoneme*

GMM *Gaussian Mixture Model*

HMM *Hidden Markov Model*

HRI *Human Robot Interaction*

HTML *HyperText Markup Language*

HTTP *HyperText Transfer Protocol*

ICAD *International Conference on Auditory Displays*

ICSI *International Computer Science Institute*

ICT *Information and Communication Technology*

IPC *Inter Process Communication*

IR *Information Retrieval*

IU *Image Understanding*

IUI *Intelligent User Interfaces*

LM *Language Model*

LP *Linear Prediction*

LSA *Latent Semantic Analysis*

MAP *Maximum A Posteriori*

MFCC *Mel-Frequency Cepstral Coefficients*

MI *Multimodal Interaction*

MT *Machine Translation*

NII *National Institute of Informatics*

NLG *Natural Language Generation*

NLP *Natural Language Processing*

NLU *Natural Language Understanding*

NP *Non-deterministic Polynomial time*

PCA *Principal Component Analysis*

PCM *Pulse Code Modulation*

PLSA *Probabilistic Latent Semantic Analysis*

POS *Part-Of-Speech*

SALT *Speech Application Language Tags*

SC *Speech Communication*

SDS *Spoken Dialog Systems*

SIP *Session Initiation Protocol*

SLP *Spoken Language Processing*

SLU *Spoken Language Understanding*

SSI *Silent Speech Interfaces*

STSA *Short-Time Spectral Amplitude*

SV *Speaker Verification*

SVM *Support Vector Machine*

SVS *Speaker Verification Spoofing*

TTS *Text-to-Speech Synthesis*

UBM *Universal Background Model*

UI *User Interfaces*

VoiceXML *Voice eXtended Markup Language*

VUI *Voice User Interfaces*

VR *Virtual Reality*

WER *Word Error Rate*

XML *eXtended Markup Language*

1 Speech Communication

1.1 Introduction

This habilitation thesis covers the scientific field of *Speech Communication* (SC). This field is divided into several sub-areas according to the division from the 2015 Interspeech conference [ISCA, 2015], which is the largest conference in the field of *Speech Communication*. At Interspeech we can find 12 scientific areas where I will show my contributions in 10 of these scientific areas ("Area 1. Speech Perception, Production and Acquisition", "Area 2. Phonetics, Phonology, and Prosody", "Area 3. Analysis of Paralinguistics in Speech and Language", "Area 4. Speaker and Language Identification", "Area 5. Analysis of Speech and Audio Signals", "Area 7. Speech Synthesis and Spoken Language Generation", "Area 8. Speech Recognition – Signal Processing, Acoustic Modeling, Robustness, Adaptation", "Area 9. Speech Recognition – Architecture, Search, and Linguistic Components", "Area 10. Speech Recognition – Technologies and Systems for New Applications", "Area 11. Spoken Language Processing – Dialog, Summarization, Understanding"). I made no contributions to two areas, namely "Area 6. Speech Coding and Enhancement" and "Area 12. Spoken Language Processing – Translation, Information Retrieval, Resources", which will also be described briefly.

There is hardly any textbook that fully covers the whole field of *Speech Communication*. Good introductions into a large part of SC are given in Benesty et al. [2007]; O'Shaughnessy [1987] and Huang et al. [2001].

Speech Communication has played an important role in AI from its beginnings [Nilsson, 2010]. From early research on neural networks in 1957 where perceptrons were seen as "potential models of human learning, cognition, and memory" [Nilsson, 2010][p. 64] to early *Natural Language Processing* (NLP) systems [Nilsson, 2010][p. 103] and first attempts in *Natural Language Understanding* (NLU) and *Automatic Speech Recognition* (ASR) in 1970 [Nilsson, 2010][p. 211] the problem of building natural speech and language based user interfaces was at the core of AI research.

If we define AI as the field that aims to build machines that are able to act humanly [Russell and Norvig, 1995][p. 5] and are thereby able to pass the Turing test [Turing, 1950] *Speech Communication* is necessary to realize such intelligent machines. We also may need such a component if we define AI more generally as the field that aims to build intelligent agents that are able to act rationally, which is the definition used in Russell and Norvig [1995][p. 7].

Since these first attempts we have seen an enormous progress in *Speech Communication*, which has led to the deployment of several consumer products. Most recent developments that happened in the last years are Apple's personal digital assistant SIRI (a multimodal dialog system) [Bellegarda, 2014] and Google glass [Google, 2015] that includes voice commands and touch input, which can also be combined. Recent progress in virtual reality shows further potential for the use of *User Interfaces* based on *Speech Communication*.

The main advantages of *Speech Communication* based *User Interfaces* (UI)s are that it is possible to develop interfaces that allow for *natural and intelligent interaction*, and interfaces that are *robust to ambiguous input*. The ultimate goal is to act according to the intentions of the user, not only react to the user input [Young, 2010]. Since we are dealing with technical interfaces we can also improve on weaknesses of human cognitive abilities. This improvement can possibly conflict with the goal of *natural interaction*.

There are several limitations of human sensory processing, which can be overcome by *Speech Communication*. Our sensory organs are perceptually limited (e.g. hearing range in 20-20.000 Hz, visible spectrum of 390-750 nm wavelength). Furthermore human working memory has a limited capacity of around seven items on average (digits, words, etc.). Human long-term memory is exposed to forgetting and retrieval problems. Furthermore our concentration is limited and influenced by external factors. All these limiting factors are irrelevant for technical user interfaces.

In mobile interfaces there are several additional advantages of *Speech Communication*. In mobile usage situations we often have *small screens in term of size and resolution*, *cumbersome text input*, or *no access to the visual display* [Niklfeld et al., 2001a]. The first two problems of classical interfaces may be overcome by today's mobile devices, but the third problem that requests a possibility for hands-free usage is still highly relevant. By using speech or multimodal interfaces we can solve this problem.

It has been shown that statistical models that can be learned from data are reliable modeling paradigms for *Speech Communication* that can reach the above-mentioned goals. Historically there has been an alternation and fusion between statistical and symbolic approaches as shown in the historical development of *Spoken Language Processing* [Jurafsky and Martin, 2000]. Jurafsky concedes a period of foundational insights in the 1940s and 1950s with formal language theory and the McCullough-Pitts neuron. This period is followed by a symbolic and a stochastic camp (1957-1970). Between 1970 and 1983 four paradigms are dominating the field, namely stochastic modeling, logic-based modeling, NLU, and discourse modeling. From 1983 to 1993 we see an empiricist period where finite state and probabilistic models are favored. From 1994 on Jurafsky observes a coming together of the statistical and symbolic approaches.

The alternation and rediscovery of modeling approaches can also be observed by the recent rise of *Deep Neural Networks* (DNN) [Hinton et al., 2012; Zen et al., 2013] in *Speech Communication*. These successors of the McCullough-Pitts neuron exploit the

availability of large amounts of data and improvements in computing power and modeling algorithms.

The alternation between different modeling approaches also shows that the problems of developing *Speech Communication* based user interfaces are not solved by one paradigm. This also shows the inherent difficulty of these interfaces, which is true of AI in general. A problem can be called AI-complete if all other problems in AI can be reduced to this very problem by some reasonably complex procedure. AI-completeness is defined in analogy to *Non-deterministic Polynomial time* (NP)-completeness [Cook, 1971]. The main difference between AI and NP-completeness is that algorithmic solutions to NP-complete problems like "satisfiability of a formula in classical propositional logic" are known but not efficient, but to this day there is no known solution to any AI-complete problem. The class of AI complete problems must at least contain the hardest problems in AI like *Natural Language Understanding* (NLU) and *Image Understanding* (IU). We believe that also the problem of generation of natural speech output is AI-complete if we understand it as the problem of generating conversational speech with all it's features.

1.1.1 Own contributions

My own contributions to the field of SC will be given in detail below. A general overview of SC is given in my course on *Cognitive User Interfaces* (CUI) that I hold at Vienna University of Technology regularly since the winter term 2011 [Pucher, 2015b]. *Cognitive User Interfaces* (CUI) are *User Interfaces* (UI) that have the ability "to support *reasoning and inference*", "*plan* under uncertainty", "*adapt* online to changing circumstances", and "*learn* from experience" [Young, 2010]. The kind of *reasoning and inference* that is applied in different interfaces might by very different, and range from logical reasoning in *Natural Language Understanding* (NLU) [Allen, 1995] and *Image Understanding* (IU) [Crevier and Lepage, 1997] to probabilistic reasoning in statistical parametric *Text-to-Speech Synthesis* (TTS) [Tokuda et al., 2000a].

1.2 Speech Perception, Production and Acquisition

1.2.1 Overview

Studies on speech perception, production and acquisition have played an important role in the field of SC from the beginning [O'Shaughnessy, 1987; Quatieri, 2002]. These two textbooks also provide a good introduction into speech perception, production and acquisition from a *Speech Communication* point of view.

To develop human machine interfaces based on *Speech Communication* (SC) it is illustrative to investigate how humans perform the task that shall be performed by the technical interface. However, it is not always necessarily the best path to try to learn to fly from birds, but often beneficial to search for a more technically minded solution. In

TTS for example the state-of-the-art method of unit selection uses dynamic programming to perform a search on a large speech unit database [Hunt and Black, 1996], which is surely not the way in which humans produce speech. In many areas of SC large amounts of training data are used, which also make it impossible to relate these models directly to human processing. In language modeling for speech recognition for example we use massive amounts of text data to train n-gram models [Jelinek, 1976]. The same is true these days for speech synthesis [Zen et al., 2013] and acoustic modeling for speech recognition [Hinton et al., 2012].

Important current fields of research within this topic are L1 (first language) / L2 (second language) perception [Ooigawa, 2015], production [Grohe et al., 2015] and acquisition [Ordin and Polyanskaya, 2015], brain models for perception [Dehaene-Lambertz et al., 2002], production [Amunts et al., 1999] and acquisition [Kuhl, 2004], and new techniques for speech production measurement and analyses [Andrade-Miranda et al., 2015; Csapó and Lulich, 2015].

1.2.2 Own contributions

Interaction speech production-speech perception

In a recent paper [Pucher et al., 2015] we have shown that listening to one's own synthetic voice increases engagement and performance of blind school children in audio games. For this evaluation we developed an audio-only labyrinth game to measure engagement time and an audio-only memory game to measure performance. Familiar voices like teacher's voices show a trend of increased engagement and performance, but more experiments are needed for verifying this hypothesis. For blind users that are using speech synthesis on a regular basis there is a need to make their synthesizer experience more engaging and pleasurable, which can be accomplished by using their own or familiar voices in the synthesizer.

1.3 Phonetics, Phonology, and Prosody

1.3.1 Overview

An introduction into the field of phonetics is given in Ladefoged and Johnson [2014], which distinguishes between articulatory phonetics, acoustic phonetics, and auditory or perceptual phonetics. Stevens [1998] provides an introduction to acoustic phonetics, auditory phonetics is covered in Johnson [2011], and articulatory phonetics in Gick et al. [2012]. Good introductions to phonology can be found in Clark and Yallop [2000]; Hall [2011]; Spencer [1995]. Prosody is mostly dealt with in the context of phonology.

Articulatory phonetics, acoustic phonetics, and auditory or perceptual phonetics have all different applications in *Speech Communication*. Phonology is concerned with the

systematic view of the organization of patterns of sounds in languages, and prosody is concerned with higher level units like syllables and their intonation, duration, and phrasing patterns that are used to convey meaning or are used for specific styles of speaking.

The importance of phonetics for speech processing is sometimes forgotten when we deal with standard languages where a large amount of phonetic knowledge is already encoded in the respective corpora. But for new languages and dialects we need a detailed phonetic analysis of the respective language variety. For building a state-of-the-art speech synthesis system for a new dialect for example we need to define the phonetic inventory, phonetic classes, and construct a corpus that covers many phonetic features.

1.3.2 Own contributions

Sociophonetics

We showed how dialects for which neither a corpus nor a sufficient linguistic description exist are modeled and synthesized on the basis of a comparatively small corpus (600 utterances per dialect) [Pucher et al., 2012b]. In this publication we focused on the quality of the recordings, the methods of data collection, the quality of the speech corpus for the synthesis, the characteristics of the underlying phone set, and on the analysis of the material.

We also showed how to realize a variety-slider on basis of our developed algorithms [Pucher et al., 2012c]. The realization of a variety slider is a prerequisite to synthesize different aspects of varieties. This paper was presented at a conference for dialectology and was aiming to describe the developed algorithms in a more linguistic setting.

I was also supervising a diploma thesis at the University of Saarbrücken that was carrying out a quantitative and phonetic analysis of non-linguistic particles in spontaneous speech of Viennese sociolects [Bruss, 2008].

1.4 Analysis of Paralinguistics in Speech and Language

1.4.1 Overview

The recently published book by Schuller and Batliner [2013] gives a very good introduction into the rising field of paralinguistics from a computational perspective.

Paralinguistics can be defined as "the discipline dealing with those phenomena that are modulated onto or embedded into the verbal message, be this in acoustics (verbal, non-verbal phenomena) or in linguistics (connotations of single units or of bunches of units." [Schuller et al., 2013]. The field of paralinguistics was first described by Trager

[1958]. Language traits that are dealt with in analysis of paralinguistics can be long term (biological, personal, group traits), medium term (health state, attitude), or short-term states (emotion, laughter, sighs) [Schuller et al., 2013]. Sociolect and dialect are sometimes also subsumed under the long-term traits [Schuller et al., 2013], whereas we would subsume these phenomena under the field of sociolinguistics or sociophonetics. From a speech processing point of view however it might be useful to keep this long, medium, and short-term classification.

Recent investigations in the field of analysis of paralinguistics dealt with the detection of degree of nativeness based on acoustic features, degree of Parkinson's condition based on speech analysis, and the eating condition [Schuller et al., 2015]. In the field of speech synthesis paralinguistic phenomena have been mainly studied in the context of conversational speech synthesis [Campbell, 2006; Urbain et al., 2013].

1.4.2 Own contributions

Non-verbal communication

I was a national Austrian representative of the EU-COST action 2102, which dealt with cross-modal analysis of verbal and non-verbal communication [Esposito, 2006]. In 2005 I was organizing a workshop on " Combining Speech and Sound in the User Interface" at the *International Conference on Auditory Displays* (ICAD) [Fröhlich and Pucher, 2005]. At this workshop we brought together auditory display and speech-based UI designers to discuss topics relevant for both communities. We had a range of presentations from dealing with the effects of speech and non-speech sounds on short-term memory and possible implications for in-vehicle use [Vilimek and Hempel, 2005] to the realization of conversational speech synthesis systems [Massimino, 2005].

1.5 Speaker and Language Identification

1.5.1 Overview

A recent introduction into speaker recognition (identification and verification) is given by Beigi [2011]. Introductory and advanced chapters on speaker recognition can also be found in Lee et al. [2012]. O'Shaughnessy [1987] also contains a separate chapter on speaker recognition. Automatic language identification is introduced in the journal paper by Zissman and Berkling [2001].

In speaker identification we aim at the identification of a speaker given a speech sample. In speaker verification we have a given identity claim, that a speech sample was spoken by a certain speaker, and our task is to verify or falsify that claim. Language identification aims to identify a language from given speech input.

1.5.2 Own contributions

Automatic Speaker Verification Spoofing and Countermeasures

In [De Leon et al., 2012] we evaluated the security of speaker verification concerning spoofing attacks from an adaptive *Hidden Markov Model* (HMM)-based speech synthesis system and developed methods for the detection of synthetic speech, which are based on relative phase shift features. This journal paper is also a principal reference of this habilitation thesis. In a previous conference paper [De Leon et al., 2011] we published a first version of this evaluation. In an earlier paper [De Leon et al., 2010b] we investigated different synthetic speech detection methods that were proposed in the literature and showed that they are not effective for HMM synthesized speech. We also evaluated a new detection measure based on speech recognition error rate. In the first paper on this topic [De Leon et al., 2011] we evaluated the speaker verification systems using a small Austrian German speech database.

1.6 Analysis of Speech and Audio Signals

1.6.1 Overview

This field of *Speech Communication* deals with speech acoustics and the basic methods and algorithms for speech analysis and representation as well as speech segmentation, pitch and harmonic analysis and voice activity detection. Introductions into the basic algorithms for speech representation are given in Quatieri [2002] and O'Shaughnessy [1987].

Recent topics in the analysis of speech and audio signals have been the development of robust methods for voice activity detection [Sriskandaraja et al., 2015] and acoustic indoor localization [Choi et al., 2015], the application of speech processing methods to non-speech audio signals like vocalizations of birds [O'Reilly et al., 2015] and non-speech audio events [Phan et al., 2015], speech processing in the phase domain [Mowlaee et al., 2014; Loweimi et al., 2015], as well as the application of new concepts like DNNs [Espi et al., 2015].

1.6.2 Own contributions

Audio signal analysis and representation

In the topic of speech processing in the phase domain we investigated the importance of the phase information in the perceptual quality of speech signals [Saratxaga et al., 2012]. Many speech synthesizers do not use the original phase information of the signals assuming their contribution is almost inaudible. We have evaluated the perceptual

impairments produced in speech signals when their original phase information is disregarded and substituted by different approximations. Our results showed that these manipulations produce audible degradation of the speech signal, thus suggesting that signal quality can be improved using more elaborate phase models.

Pitch and harmonic analysis, Singing analysis

Within our project on the analysis and statistical modeling of opera singing [Pucher and Yamagishi, 2014] we evaluated an efficient wide-range pitch estimation algorithm based on spectral correlation [Villavicencio et al., 2015]. We showed that the algorithm achieves high performance on wide pitch-range signals independently of the voice height. We compared the technique with three state-of-the-art techniques on natural opera singing observing rich pitch content.

1.7 Speech Coding and Enhancement

1.7.1 Overview

An overview of speech coding can be found in Chu [2004]. Introductions to speech coding are also given in Quatieri [2002] and O'Shaughnessy [1987] and in the classical textbook on digital processing of speech signals by Rabiner and Schafer [1978].

In speech coding the speech signal is converted and compressed into a sequence of bits for storage and transmission [O'Shaughnessy, 1987]. Speech coders can be divided into waveform approximating coders, parametric coders, and hybrid coders [Kondoz, 2005]. Examples of waveform approximating coders are *Pulse Code Modulation* (PCM) and *Code Excited Linear Prediction* (CELP). Examples of parametric coders are *Linear Prediction* (LP) based coders and harmonic or sinusoidal coders where the speech signal is modeled by a combination of time varying sinusoidal or harmonic signals. Parametric coders are also very important for speech synthesis. The *Adaptive Multi-Rate* (AMR) audio codec that is widely used in telecommunications is an important example of a hybrid coder.

Current research topics in speech coding are contributions to the *3rd Generation Partnership Project* (3GPP) codec for *Enhanced Voice Services* (EVS) [Dietz et al., 2015; Disch et al., 2015; Eksler et al., 2015; Atti et al., 2015; Jokinen et al., 2015; Nagisetty et al., 2015] or the use of neural network phonological features for low bit rate speech coding [Cernak et al., 2015; Asaei et al., 2015].

In speech enhancement a degraded speech signal is processed to enhance it. The degrading can come from a noisy communication channel or environment noise. Important

methods for *Short-Time Spectral Amplitude* (STSA) based speech enhancement are spectral subtraction, maximum likelihood based estimation, and Wiener filtering [Kondoz, 2005].

Current research topics in speech enhancement are the modeling of temporal dependencies for robust parameter estimation [Wong et al., 2015], and the usage of DNN based learning algorithms for separation [Zhang and Wang, 2015], echo suppression [Lee et al., 2015], or text-informed speech enhancement [Kinoshita et al., 2015].

1.7.2 Own contributions

Until now I have not been working in the area of Speech Coding and Enhancement and have therefore no own contributions in this area.

1.8 Speech Synthesis and Spoken Language Generation

1.8.1 Overview

A recent introduction into speech synthesis is given in Taylor [2009]. This textbook covers unit selection and HMM-based methods. A good recent textbook in German covering concatenative synthesis is Pfister and Kaufmann [2008]. Visual speech synthesis is introduced in Parke and Waters [2008] from a facial animation point of view, and covered in several chapters in Bailly et al. [2012].

In *Text-to-Speech Synthesis* (TTS) we are given a sequence of words, and want to generate an acoustic speech signal of a specific speaker. This can be extended to concept-to-speech synthesis where we convert a given concept into an acoustic signal. A concept is a semantic representation of the meaning of an utterance, e.g. a certain command, wish, statement. Concept-to-speech synthesis can be split into the two separate problems of *Natural Language Generation* (NLG) and *Text-to-Speech Synthesis* (TTS).

In the last decade we can see four main speech synthesis methods that were developed. Unit-selection [Hunt and Black, 1996], HMM-based [Tokuda et al., 2000b], hybrid [Ling and Wang, 2006], and DNN-based [Zen et al., 2013] systems. While unit-selection was already invented in the 90s the other three methods were more recently introduced, although first attempts to use neural networks in speech synthesis were already made in the 90s [Riedi, 1995; Traber, 1991] but with a more limited scope than current attempts. With these technologies we can produce intelligible, natural, and flexible speech synthesis systems. A problem that is still unsolved is the creation of a system that speaks like in a natural human-human conversation in any speaker's voice. This requires the solution of problems like variety switching [Toman et al., 2015], prosody modeling, modeling of non-linguistic particles (filled pauses, hesitations, laughing, whispering) [Campbell, 2006] as well as the ability to model many large context dependencies.

The four mentioned methods need large amounts of high quality recordings (> 32 kHz sampling rate, recorded in sound booth) as training data for achieving high quality synthesis. The fact that we are able to process these large amounts of speech data today is a result of progress in algorithms, processing, and data storage power. The continuing increase of available data and processing power will surely lead to further progress in speech synthesis. Additionally we however need to deepen our understanding of the speech production and learning process that allows humans to learn speech from a limited amount of multimodal input data.

The rising quality of these speech synthesis systems and the need for speech output interfaces has led to a number of new applications where synthesis technology is used. Speech synthesis is used in web readers, screen readers for blind users [Pucher et al., 2015, 2010a], spoken dialog systems for call center automation and information systems, car navigation systems, and personal digital assistants [Bellegarda, 2014].

The field of visual speech synthesis is fairly well established and a variety of approaches have been developed since the first rule based systems [Cohen and Massaro, 1993]. Video-based systems [Bregler et al., 1997; Ezzat et al., 2002] and other data-driven approaches [Bailly et al., 2003; Deng and Neumann, 2006; Theobald et al., 2004] have been developed.

Visual speech can be characterized by the movements of marker points in the face. Phonemes can be grouped into phoneme equivalence classes or visemes to cover visual speech characteristics. The consonants /p/, /b/, and /m/ for example correspond to the same viseme since they are visually equivalent, but acoustically very different.

One reason why visual speech synthesis is difficult, is the high sensitivity of human viewers to any kind of errors, since humans are trained very well to recognize facial movement [Berger et al., 2011]. This is also related to the uncanny valley paradox [Mori et al., 2012], which states that beyond a certain point of realism a virtual character gets less accepted if it becomes more realistic. This paradox arises, because humans are able to spot small errors in a very realistic virtual character, which makes the character seem creepy.

1.8.2 Own contributions

In 2015 I was area chair for “Speech Synthesis and Spoken Language Generation” at the INTERSPEECH conference, the largest and most important conference in the field of *Speech Communication* (SC).

In the summer term of 2008 I held a seminar on “Speech Synthesis” at the Signal Processing and Speech Communication Laboratory (SPSC Lab) at Graz University of Technology. According to the technologies that were most relevant at that time the contents where mainly unit-selection and HMM-based systems.

Unit selection speech synthesis

Together with the Scottish company Cereproc I was involved in the development of "Leopold" the first synthetic voice for Austrian German in 2010, which was integrated into a web reading service for the Website of the City of Vienna. The voice can also be bought from the web store for different platforms [Cereproc, 2010].

Between 2007 and 2009 I was principal investigator of the project "VSDS - Viennese Sociolect and Dialect Synthesis" [Pucher, 2007b] that was funded by the Vienna Science and Technology Fund (WWTF). Within this project we investigated the modeling of Viennese varieties with unit-selection and HMM-based methods. During this project I was also visiting the Centre for Speech Technology Research (CSTR) in Edinburgh to start my collaboration with Junichi Yamagishi (now at NII, Japan and CSTR, Edinburgh) on statistical parametric synthesis.

In Pucher et al. [2010b] we showed how to optimize the phonetic encoding for Viennese dialect unit selection speech synthesis. We showed how to find optimal different phone sets for the tasks of automatic alignment and *Grapheme-To-Phoneme* (G2P) conversion.

In Kranzler et al. [2009] we showed how to develop a text-to-speech engine with an Austrian German corpus. The results of this paper were also the basis of a master thesis at Graz University of Technology [Kranzler, 2008] that was co-supervised by me. In an earlier paper [Neubarth et al., 2008] we developed methods for the modeling of Austrian dialect varieties for TTS.

Within a project that was carried out together with company partners at FTW we developed methods and a system for combining non-uniform unit selection with diphone based synthesis [Pucher et al., 2003a]. With this system a user could add an additional prompt to a database of existing speech recordings and optimize the prompt to improve naturalness and emotionality.

Statistical parametric speech synthesis

I was principal investigator of two research projects that were dealing with developing new methods of HMM-based speech synthesis for modeling of language varieties. The project "AMTV - Acoustic modeling and transformation of varieties for speech synthesis", from 2012 to 2016 [Pucher, 2012a] and the project "SALB - Speech synthesis of auditory lecture books for blind children" from 2013 to 2015 [Pucher, 2013]. The first project was funded by the Austrian Science Fund (FWF) and dealt with variety modeling and transformation. The second project was funded by the Austrian ministry of Science and Research (BMWF) within the Sparkling Science program where research institutions are working together with schools.

On our work on HMM-based synthesis for language varieties I held two invited talks, one at the *National Institute of Informatics* (NII) in Tokyo, Japan on "Interpolation of language varieties in HMM-based speech synthesis", and one at a Dagstuhl workshop that was dedicated to multilinguality in speech research on "Acoustic modeling, interpolation, and transformation of language varieties for speech synthesis".

Recently we developed [Toman et al., 2015] a method for unsupervised interpolation of language varieties for speech synthesis. This journal publication is also one of the principal references of this habilitation thesis. With this method we can create in-between varieties from a standard and a variety (dialect, sociolect, accent) model in an unsupervised way. This work is also an important part of Markus Toman's PhD thesis [Toman, 2016], which I was co-supervising at FTW.

In Pucher et al. [2015c] we showed how the adaptive HMM-based modeling approach can be applied for the modeling of the Albanian dialects Tosk and Gheg. For this application we developed a front-end for Albanian G2P together with our partners from University of Prishtina.

Within the Pucher [2012a] project we also developed methods for the modeling a language varieties that are able to exploit the overlapping linguistic and acoustic information between these varieties. We applied multi-variety adaptive acoustic modeling in HSMM-based speech synthesis [Toman et al., 2013c] and showed how to select phone sets for HMM-based dialect speech synthesis [Pucher et al., 2011].

Within the Pucher [2007b] project we started our work on statistical parametric speech synthesis with the development of a method for the supervised interpolation of dialects, showcasing the method for generating in-between varieties of Viennese and Standard Austrian German [Pucher et al., 2010b]. This journal publication is also a principal reference of this habilitation thesis. Parts of this work were carried out within a master thesis at Vienna University of Technology that was co-supervised by me [Schabus, 2009] and that won the thesis award of the Austrian Computer Society (OCG).

Prosody modeling and generation

In the project Pucher [2013] we worked on the topic of prosody modeling to generate high-quality and intelligible fast speech. The production of fast speech is important for blind users that use fast speech output for compressed information presentation. In a recent journal paper [Valentini-Botinhao et al., 2015] and a previous conference paper [Valentini-Botinhao et al., 2014] we compared linear and non-linear methods for speech compression and showed that the non-linear methods outperform the linear ones depending on the intelligibility of the fast speech data. In earlier experiments we evaluated different non-linear methods and showed that the best performing method was the one that only interpolated between duration models, and used spectrum and F0 models from the normal speaking rate data [Pucher et al., 2010a]. All three papers are principal reference of this thesis.

Expression, emotion and personality generation

The speech synthesis output component is an important part of the persona of a speech based user interface. When we started to work on speech synthesis of dialects we investigated which application scenarios are appropriate for dialect voices [Pucher et al., 2008]. For this investigation we performed a user workshop and conducted listening experiments to evaluate which features can be recognized in dialect voices. In Pucher et al. [2012d] we showed a framework for creating regionalized virtual avatars by using adaptive audio-visual dialect speech synthesis.

Synthesis of singing voices

Since 2013 I am collaborating with colleagues from Japan on the topic of synthesis of opera singing [Pucher and Yamagishi, 2014]. Since then we have performed high quality recordings of professional opera singers in Vienna, and have built a state-of-the-art German opera singing synthesis system based on these recordings. Within this project I was also visiting NII in Tokyo in 2014 and working on a master thesis in computer science at Vienna University of Technology [Pucher, 2015a]. We have been working on the optimization of pitch extraction methods for opera singing [Villavicencio et al., 2015] and have also evaluated operatic singing synthesis of a mezzo-soprano and bass voice using different vibrato modeling methods [Pucher et al., 2016a].

Voice modification, conversion and morphing

As part of the Pucher [2012a] project we have developed several conversion methods in the context of language varieties and visual speech. In Toman and Pucher [2015a] we showed how to transform the acoustic model of a second language learning (L2) accented voice into an accent-free voice of the same speaker, which is a useful application for language learning. In previous work we showed how to transform acoustic models between standard and dialect while retaining speaker similarity [Toman and Pucher, 2013; Toman et al., 2013b]

Avatars and talking faces

Between 2011 to 2014 I was principal investigator of the project “AVDS - Adaptive Audio-Visual Dialect Synthesis” [Pucher, 2011] funded by the Austrian Science Fund (FWF). In this project we investigated statistical approaches for audio-visual modeling. I was local main organizer of the conference “FAAVSP 2015 - The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing” [Pucher et al., 2015a,b; Pucher, 2015d] and the conference “FAA 2012 - The ACM 3rd International Symposium on Facial Analysis and Animation” [Pucher et al., 2012a; Pucher, 2012b] in

Vienna. In July 2011 I held a seminar on Audio-Visual Speech Synthesis at the Signal Processing Laboratory (Aholab) of the University of the Basque Country. The contents of the seminar were HMM-based systems and their application for audio-visual speech synthesis.

In the field of audio-visual synthesis we showed that joint modeling of acoustic and visual features within a statistical parametric system can improve the audio-visual synthesis quality compared to single modality modeling [Schabus et al., 2014]. This journal paper was also an important part of Dietmar Schbaus' PhD thesis at Graz University of Technology [Schabus, 2014] that I was co-supervising and is also one of the principal references of this habilitation thesis.

In Hollenstein et al. [2013] we showed that the acoustic output of a synthesizer can be controlled by the visual features that are projected in a low dimensional space to allow for controlling with one or two parameters. This conference paper is also a principal reference of this habilitation thesis and was also part of a master thesis at Vienna University of Technology that was co-supervised by me [Hollenstein, 2013].

In a series of paper we were working on additional aspects of audio-visual modeling in the statistical framework. In Schabus and Pucher [2015] we evaluated the use of visual dialect data for synthesis and showed that a phone mapping can achieve similar quality to audio-visual training data. Previously we evaluated [Schabus et al., 2013] in an objective and subjective evaluation which feature set size we should use for audio-visual speech synthesis. We also applied the adaptive HMM-based framework to audio-visual modeling [Schabus et al., 2012] and showed that adaptive models can lead to improved quality when small amounts of training data are available. This conference paper is also a principal reference of this habilitation. In the first paper were we applied the HMM-based framework to visual modeling [Schabus et al., 2011] we showed that it is feasible to produce both acoustic speech parameters and animation parameters by a maximum likelihood parameter generation algorithm from models that were trained on such a synchronous corpus.

Tools and data for speech synthesis

From the various research projects we contributed several tools and data sets to the research community. We have developed and released an HMM-based Austrian German open source voice for the Festival [Festival, 2015] speech synthesis system [Toman et al., 2013a] and our own lightweight speech synthesis framework [Toman, 2013; Toman and Pucher, 2015b], which is also part of the extensions to the HMM-based speech synthesis system HTS [HTS, 2015].

We also developed and released 3 Viennese sociolect open source unit selection voices for the Festival speech synthesis system [Pucher et al., 2010c].

In the audio-visual domain we released the Bad Goisern and Innervillgraten Audio-Visual Dialect Speech Corpus (GIDS) [Schabus and Pucher, 2014a] and a triple-modality corpus containing articulatory, visual and acoustic data of a speaker (Multi-Modal Annotated Synchronous Corpus of Speech (MMASCS)) [Schabus et al., 2014; Schabus and Pucher, 2014b]. In Schabus et al. [2012] we also described how to build a synchronous corpus of acoustic and 3D facial marker data that is suitable for adaptive audio-visual speech synthesis.

Evaluation of speech synthesis

One of my very early papers was dedicated to the evaluation of speech synthesis systems on mobile platforms [Pucher and Fröhlich, 2005]. In that paper we evaluated the quality and naturalness of different speech synthesis methods and speech codecs on mobile platforms.

1.9 Speech Recognition – Signal Processing, Acoustic Modeling, Robustness, Adaptation

1.9.1 Overview

A recent textbook in German on speech recognition with HMMs is Pfister and Kaufmann [2008]. A classic introduction into this topic is Rabiner and Juang [1993]. Further good textbooks on speech recognition with HMMs are Huang et al. [2001] and Furui [2000]. Jurafsky and Martin [2000] also provides a good introduction into speech recognition with HMMs from a more linguistic angle. A very recent textbook that also covers DNNs for acoustic modeling is Yu and Deng [2014].

In *Automatic Speech Recognition* (ASR) we aim at the recognition of a word sequence from an acoustic signal. In the last decades there has been significant progress in speech recognition, both in training the *Acoustic Model* (AM) where we aim at modeling the probability of an acoustic signal given a word sequence, and in *Language Model* (LM) training where the probability of word sequences is modeled [1].

In recent years the standard architecture for ASR, consisting of HMMs as acoustic models combined with n-grams for language modeling [Jelinek, 1976; Baker, 1975; Bahl et al., 1983], have been replaced or extended by architectures using deep learning methods [Hinton et al., 2012] mainly for acoustic modeling. These methods can and need to use large amounts of training data, which are available today. Deep learning methods such as recurrent neural networks have already been applied in ASR some time ago [Morgan and Scofield, 1991], but as in other technical fields the availability of more data, better

[1] Language modeling is part of the next topic "Speech Recognition – Architecture, Search, and Linguistic Components"

algorithms, and faster hardware partly explains the comeback of deep learning methods.

Important challenges for speech recognition are still the adaptation to changing environments coming from different background noises or speaker characteristics such as accents, dialects, and speaking styles.

1.9.2 Own contributions

Acoustic modeling for conversational speech (dialog, interaction)

In acoustic modeling we published one paper where we showed how phonetic distance measures can be applied to optimize a speech recognizers vocabulary and grammar [Pucher et al., 2007]. The investigated phonetic distance measures were based on the minimum-edit-distances between phonetic transcriptions and the distances between HMMs.

1.10 Speech Recognition – Architecture, Search, and Linguistic Components

1.10.1 Overview

Language modeling and system development is covered in the textbook Huang et al. [2001]. Language modeling from an NLP perspective is covered in Manning and Schütze [1999]. A very good introduction into language modeling topics for ASR is found in Jelinek [1997].

There are four prominent groups of language modeling techniques. These are n-gram language models [Jelinek, 1976; Baker, 1975; Bahl et al., 1983], structured language models [Chelba et al., 1997; Chelba and Jelinek, 1998; Wang et al., 2005], topic models [Bellegarda, 2000; Gildea and Hofmann, 1999] and WordNet-based models [Budanitsky and Hirst, 2001; Demetriou et al., 1997][Pucher, 2005, 2007c].

Conditional n-gram models can be conditioned on different types of events. Word-token-based n-gram models and class-based n-gram models [Heeman, 1998] are very prominent examples. Structured language models can focus on syntactic or semantic structure. These models can provide an integration of *Automatic Speech Recognition* (ASR) and written language understanding which results in a full approach for *Spoken Language Understanding* (SLU).

Topic models can be either based on *Latent Semantic Analysis* (LSA) or *Probabilistic Latent Semantic Analysis* (PLSA) models. The concept of LSA was originally used in *Information Retrieval* (IR) to define similarities between queries and documents and was then applied for language modeling. PLSA models derive the probabilities between a

word and a document or history of words directly, without estimating a similarity first. The basic idea underlying these models is to introduce a latent topic variable. Models that are based on WordNet can either be graph-based or text-based.

Recently DNN-based language models have also been investigated [Arisoy et al., 2012], although word-based n-gram models still remain very prominent.

1.10.2 Own contributions

Language modeling for conversational speech (dialog, interaction)

During my PhD [Pucher, 2007a] I was visiting the *International Computer Science Institute* (ICSI) in Berkeley in 2005 within the visiting program of the European Union 6-th FP IST Integrated Project *Augmented Multi-Party Interaction* (AMI). There I also developed an LSA Language Modeling Toolkit for training and testing LSA models with n-gram models. This toolkit is based on SRILM and svdpackc.

In the area of WorNet-based language models we investigated how WordNet-based models can be combined with word-based n-gram models that model word order [Pucher, 2007c]. Previously we evaluated the performance of different WordNet similarities for word prediction [Pucher, 2005] and showed that a combination of similarities achieves the best performance depending on the *Part-Of-Speech* (POS) classes of words.

In LSA-based modeling we showed how to apply the LSA-based models for meeting recognition [Pucher and Huang, 2005]. We also showed how to combine multiple LSA models for meeting recognition [Pucher et al., 2006a] and how to optimize this combination [Pucher et al., 2006b].

1.11 Speech Recognition – Technologies and Systems for New Applications

1.11.1 Overview

Recent new applications of speech synthesis are in the domain of multimodal systems and silent speech interfaces. Multimodal speech processing is introduced in Bailly et al. [2012]. An introduction into multimodal interfaces is provided by Oviatt [2003]. Silent speech interfaces are introduced in Denby et al. [2010].

Multimodal Interaction is of special importance for natural user interaction since all human communication is multimodal involving gesture and posture interaction. The McGurk effect further shows that visual speech information is integrated into acoustic speech perception [McGurk and MacDonald, 1976]. If we see a visual signal of a person

saying /ga-ga/ and hear an acoustic signal of a person saying /ba-ba/, then we actually hear /da-da/. The visual and acoustic signal have of course to be synchronized.

Multimodal interfaces consist of a combination of several input and output modalities like speech, vision, and gesture [Oviatt, 2003]. The combination of multiple modalities can happen at an earlier or later stage, which can be described as early or late fusion. An example of early fusion is the combination of acoustic and visual speech on a feature or state level in an audio-visual synthesis or recognition system [Schabus et al., 2014] [Dupont and Luettin, 2000]. Late fusion can be used in multimodal dialog systems to combine the output of visual and speech input into a joint recognition result [Lalanne et al., 2009] [2]. For both types of fusion methods synchronization is an important issue. It has been shown in many fields that the combination of multiple modalities can lead to more reliable and intuitive systems [Oviatt and Cohen, 2000]. This is also true for the related approach of sensor fusion where data from multiple sensors is combined, where we have for example shown how acoustic and visual sensors can be applied for tracking cars on highways [Pucher et al., 2010d].

Silent Speech Interfaces (SSI) are new applications that enable speech communication to take place when an audible acoustic signal is unavailable [Denby et al., 2010]. These systems use non-acoustic signals that are generated from the human speech production process. Applications are for users that have undergone a laryngectomy or that cannot speak due to paralysis, for privacy-aware and non-disturbing mobile speech communication, and for speech processing in noisy environments.

1.11.2 Own contributions

Multimodal systems

Within a project that was conducted together with industrial partners at FTW we investigated mobile multimodal next generation applications [FTW, 2004]. We developed a server-based platform that could render multimodal content based on *eXtended Markup Language* (XML) representations for different device types. Similar applications are today available in consumer products [Bellegarda, 2014]. For this platform we also developed two prototypical applications: a multiplayer quiz game (MONA@play) and a unified messaging application (MONA@work). These applications could be deployed on a wide range of devices including PocketPC PDAs, Symbian based smartphones and even low-end mobile phones with a WAP browser only, which are not relevant today anymore [Niklfeld et al., 2005a; Anegg et al., 2004].

In a series of papers [Niklfeld et al., 2005b, 2002c,a,b] we developed mobile multimodal user interface technologies for the then upcoming GPRS, UMTS, and WLAN networks. The focus of this work was on the application and extension of existing standards like *Voice eXtended Markup Language* (VoiceXML) and *Speech Application Language Tags*

[2] Lalanne et al. [2009] distinguishes seven levels where multimodal fusion can happen

(SALT), and the investigation on which types of multimodality are realizable with then existing terminal and network capabilities.

Previously we developed a first architecture for multimodal interfaces [Niklfeld et al., 2001c,b,a] consisting of a visual (*HyperText Markup Language* (HTML), applet) and a VoiceXML browser that communicate with an application server via *HyperText Transfer Protocol* (HTTP) and *Inter Process Communication* (IPC). One of the main challenges of this early implementation was the synchronization between visual and speech modality.

New paradigms (e.g. artic. models, silent speech interfaces, topic models)

One example for silent speech interfaces uses *ElectroMagnetic Articulography* (EMA) sensors for measuring the movement of a fixed set of points in the vocal tract (e.g. on the tongue and lips). Recently we showed how acoustic speech can be reconstructed with EMA plus visual marker recordings, and that this joint reconstruction outperforms visual or EMA-based only reconstruction [Pucher and Schabus, 2015].

1.12 Spoken Language Processing – Dialog, Summarization, Understanding

1.12.1 Overview

A good introduction into spoken dialog systems from a user interface design perspective is given in Cohen et al. [2004]. An analysis of *Spoken Dialog Systems* (SDS) from a usability perspective is provided by Hempel [2008]. *Spoken Language Understanding* (SLU) is covered in Huang et al. [2001]. Language understanding from a logic-based perspective can be found in Blackburn and Bos [2005]

A *Spoken Dialog Systems* (SDS) is a computer program that allow a user to interact with a system using speech. SDSs are the most advanced form of *Voice User Interfaces* (VUI) [Cohen et al., 2004] since they allow for full speech interaction. These systems are composed of three main components: speech synthesis for generating speech output, speech recognition for processing the acoustic input, and dialog management. Equipped with speech synthesis the dialog system is able to transform text available in written form into spoken language. Speech recognition is employed to transform spoken user utterances into written text or words using grammars. The speech recognition component can also be extended by a natural language understanding component. The dialog management component defines the interaction behavior or dialog logic of the dialog system. A standardized definition language for spoken dialog systems is VoiceXML, a markup language which is built around the web-based form-filling paradigm. The main components of a VoiceXML application are prompts (define what is said), grammars (define what can be said), and forms (define the dialog logic).

The personality or persona of a SDS must be considered , since there is no such thing as a voice user interface with no personality [Cohen et al., 2004]. The perception of sociolectal and dialectal varieties influences our evaluation of a speaker's attributes like competence, intelligence, and friendliness. The persona can be defined as a standardized mental image of a personality or character that users infer from the applications voice and language choices. Speech synthesis is an essential part of a spoken dialog system's persona.

Spoken Language Understanding (SLU) can be realized by semantically structured language models that integrate ASR and *Natural Language Understanding* (NLU) into a full approach of *Spoken Language Understanding* (SLU). The task of SLU is to find the best meaning representation given an utterance [Wang et al., 2005], the task of NLU is to find the best meaning representation given a string of words. Other approaches to NLU are logic-based approaches [Blackburn and Bos, 2005; Van Eijck and Unger, 2010] and hybrid methods [Roth and Yih, 2005; Chang et al., 2008]. Although there has been big progress in NLU, the applications of it are still restricted to certain domains [Liu et al., 2015a].

1.12.2 Own contributions

Spoken dialog systems

In the field of *Spoken Dialog Systems* we showed how Viennese language varieties can be incorporated in an SDS and thereby realize a dialog system with different personas [Pucher et al., 2010a]. In previous work [Pucher et al., 2003b] we showed how to develop an intelligent personal voice call assistant by using the *Session Initiation Protocol* (SIP) and VoiceXML.

Multimodal human-machine interaction (conversat. agents, human-robot)

Spoken Dialog Systems also play an important role in multimodal interaction. We have shown in several papers how standardized components such as SALT can be used to realize multimodal control interfaces that can be used to control a robot [Baillie et al., 2004; Pucher and Képesi, 2003]. Such a robot can be a very useful application for the home. I was also a member of EUCOG III, a European Network for the Advancement of Artificial Cognitive Systems, Interaction and Robotics [EUCOG, 2015].

Semantic analysis and classification

In one of my lectures [Pucher, 2015c] at Vienna University of Technology I teach the topic of *Computational Semantics* (CS) that deals with the extraction of semantic representations from natural language text and is a basic technology for *Spoken Language*

Understanding. In the domain of logical analysis the lecture gives an overview of Montague Semantics [Montague, 1973; Blackburn and Bos, 2005]. In the domain of representations based on similarity or probability it presents LSA, PLSA, and graph-based semantic similarity measures. In my first diploma thesis [Pucher, 2001] at University of Vienna I was presenting Post-Tarskian formal theories of truth. Tarski style semantics are the basis for logical semantics.

Evaluation of speech and multimodal dialog systems

Together with T-Labs in Berlin we were investigating automatic methods to evaluate the quality of *Spoken Dialog Systems* [Möller et al., 2008, 2007]. We evaluated phonetic similarity metrics to estimate the quality of speech recognition grammars and semantic similarity metrics for estimating the quality of the system's understanding component. This development happened within a joint project between FTW and T-Labs [FTW, 2006], where I was the project manager at FTWs side. During this project I was also visiting T-Labs in Berlin for one month to develop the evaluation system.

1.13 Spoken Language Processing – Translation, Information Retrieval, Resources

1.13.1 Overview

A good introduction to information retrieval topics can be found in Manning et al. [2008]; Manning and Schütze [1999]. A recent introduction into *Machine Translation* (MT) is provided by Koehn [2009].

Speech-to-speech translation can be considered as the holy grail of *Speech Communication* since it combines several major SC technologies like ASR, TTS, and *Machine Translation* (MT). This problem has been investigated since several decades [Waibel et al., 1991; Wahlster, 2013]. More recently specific TTS problems in speech-to-speech translation have been investigated, such as transferring the speaker identity of a speaker to the language that is translated into [Liang et al., 2010]. The core technology of MT is more located in the field of NLP and it's related conferences [Koehn et al., 2007], with speech related translation topics belonging to the field of SC. State-of-the-art methods in MT use phrase-based translation, where a statistical machine learning approach is used to train the models [Koehn et al., 2003] or parsing based translation [Li et al., 2009]. Current speech related MT topics are the adaptation of machine translation models towards ASR misrecognized speech [Ruiz et al., 2015], efficient machine translation with slow language models such as neural network or maximum entropy based language models [Emami, 2015], and the interaction of cognitive states and speech recognition performance in speech-to-speech translation systems [Akira et al., 2015].

The core methods of *Information Retrieval* (IR) also belong to the field of NLP and it's related conferences [Manning et al., 2008], with speech related *Information Retrieval* topics belonging to the field of SC. State-of-the-art methods for IR are the vector space model, language models, and different clustering methods [Manning et al., 2008]. Current speech related IR topics are in the field of information and metadata extraction from speech to improve speech based keyword search for low resource languages [Mendels et al., 2015; Zhang et al., 2015], and in the field of *Spoken Language Understanding* to improve audio-visual information retrieval [Lu et al., 2015] and spoken content retrieval [Racca and Jones, 2015].

1.13.2 Own contributions

Until now I have not been working in the area of Spoken Language Processing – Translation, Information Retrieval, Resources and have therefore no own contributions in this area.

2 Speech processing for multimodal and adaptive systems

2.1 Introduction

Adaptivity and multimodality are two important key properties for *Speech Communication* (SC) systems that are flexible, user friendly, and robust. Adaptivity and multimodality have been investigated in all types of *Speech Communication* systems and these topics are still in the research focus. This habilitation thesis shows the importance of adaptivity and multimodality in specific *Speech Communication* systems. The topic of adaptivity is investigated in *Text-to-Speech Synthesis* (TTS), *Audio-Visual Text-to-Speech Synthesis* (AVTTS), and *Speaker Verification Spoofing* (SVS), while the topic of multimodality is investigated in *Audio-Visual Text-to-Speech Synthesis* (AVTTS). Adaptivity has been important in SC for decades but today where we have to deal with big data problems and applications to many different contexts (language varieties and situations), adaptivity is even more important. Since speech is inherently multimodal, the investigation of multimodal speech allows us to make use of the full information available in the speech signal, where we show that models of different modalities can benefit from each other.

In *Text-to-Speech Synthesis* we show how to adapt an acoustic speaker model to social and regional variety [Pucher et al., 2010b], and speech duration [Pucher et al., 2010a; Valentini-Botinhao et al., 2014] and how we can apply supervised [Pucher et al., 2010b] and unsupervised variety interpolation [Toman et al., 2015] to generate in-between varieties. This interpolation allows us to develop a more flexible and personalized interaction in a spoken language interface. We also evaluate the effect of speaker familiarity on synthetic speech perception [Pucher et al., 2015].

For *Audio-Visual Text-to-Speech Synthesis* we show how a joint audio-visual modeling approach [Schabus et al., 2014], where acoustic and visual data is used jointly in the training process, can improve the audio-visual modeling and how a visual control model can be used to control the acoustic model [Hollenstein et al., 2013]. We also investigate a speaker adaptive approach [Schabus et al., 2012] that can make efficient use of audio-visual training data, which consists of synchronous 3D marker sequences of facial movement and acoustic speech recordings.

For *Speaker Verification Spoofing* we showed how adaptive speech synthesis systems can spoof speaker verification systems [De Leon et al., 2012], and how this can be blocked by adopting an adaptive speaker verification system. Since the publication of our work, several research groups have started to work on the topic of *Speaker Verification Spoofing*.

2.1.1 Text-to-Speech Synthesis

Adaptive approaches have received much attention in speech synthesis in the last decades [Tamura et al., 1998b, 2001; Yamagishi et al., 2004; Isogai et al., 2005; Yamagishi et al., 2006; Yamagishi and Kobayashi, 2007a; King et al., 2008; Yamagishi et al., 2009a] mainly due to the rise of statistical parametric speech synthesis [Zen et al., 2004]. Adaptive modeling has been applied to adapt to the speaker [Yamagishi and Kobayashi, 2007a], emotion [Qin et al., 2006], accent [Wester and Karhila, 2011; Karhila and Wester, 2011], dialect [Pucher et al., 2010b], type of articulation [Picart et al., 2014], and dysarthric speech [Veaux et al., 2012]. Interpolation of speaker models has been applied for speaker identity [Yoshimura et al., 1997], emotional speech [Tachibana et al., 2005a], speaking rate [Pucher et al., 2010a], dialect [Pucher et al., 2010b; Toman et al., 2015], and accent [Astrinaki et al., 2013]. With our work we have made significant contributions to adaptivity along several dimensions. The flexibility of *Hidden Markov Model* (HMM) based synthesis also allows for the integration of articulatory features [Ling et al., 2009] and the control of the acoustic model by articulatory features [Ling et al., 2008], formant features [Lei et al., 2011], or visual features [Hollenstein et al., 2013].

Impact of our work on dialect speech synthesis

On the adaptation and interpolation of acoustic models for speech synthesis of dialects there has not been much work before our initial publication [Pucher et al., 2010b]. On a search for "dialect speech synthesis" on Google Scholar our paper [Pucher et al., 2010b] comes up second after an IBM patent from 1997 on "Speech synthesis and analysis of dialects". There was however already much work on dialect/accent adaptation in speech recognition [Humphries et al., 1996; Kat and Fung, 1999; Huang et al., 2004], which is an important application since languages like Mandarin have many different accents and many speakers [Huang et al., 2000]. As with other topics, like the use of HMMs, it can be seen that methods or topics are first investigated in *Automatic Speech Recognition* (ASR) and then also in TTS. This can be partly explained by the larger size of the speech recognition community and the bigger market for speech recognition, which results in a larger number of proposed methods. The transfer of these ideas from recognition to synthesis does however often require the development of original methods suitable for synthesis and is not a simple application of a method to another type of problem.

Since our initial work on dialect synthesis [Pucher et al., 2010b] there have been several investigations into similar directions referencing our papers. Based on our data from Viennese speakers it could be shown that actors that did not grow up with the dialect use specific stereotypes to produce "authentic" dialect utterances [Moosmüller, 2012]. The synthetic voices that we have developed for Viennese sociolects have also been used in a social evaluation of artificial agents [Krenn et al., 2012] where it could be shown that listeners assess the artificial agents socially through the used synthetic voice.

Krenn et al. [2014] also investigated the effects of language variety and bodily behavior on the perception of a social agent along the extraversion/intraversion dimension. They could show that agents using Austrian German Standard and Viennese dialect voices were perceived as more extroverted than agents using German Standard voices. In Tamagawa et al. [2011] it was shown that the accent of the synthetic voice that is used by a robot affects the perception of the robot. When interacting with a healthcare robot people had more positive feelings towards a robot that was using a New Zealand accent, compared to US and UK accents. These results show that the synthesis of language varieties (sociolect, dialect, accents) has potential applications in the growing field of social robotics [Ge et al., 2012], a sub-field of *Human Robot Interaction* (HRI) [Goodrich and Schultz, 2007].

Navas et al. [2014] developed a TTS for the Basque dialect that is spoken in the Labourd and Lower Navarre region. Kolluru et al. [2014] developed methods for generating multiple-accent pronunciations for TTS using joint sequence model interpolation. Wutiwiwatchai et al. [2011] proposes a method for accent level adjustment in a bilingual Thai-English text-to-speech synthesis system. Such a system can be used to synthesize English words with an intermediate Thai-English pronunciation that is preferred by listeners compared to a standard English or Thai pronunciation. Ashby et al. [2010] developed a rule-based system for Luso-African varieties from Cape Verde and Mozambique for generating accent-specific phonetic transcriptions for these variants of Portuguese. Watson and Marchi [2014] describes resources that have been developed for speech synthesis of New Zealand English. We have also continued our work on dialect interpolation by extending our supervised interpolation method by an unsupervised method [Toman et al., 2015] that can generate intermediate variants fully automatic.

The methods and algorithms that we have developed are applicable for any dialect in any language, the interest in dialect speech synthesis may however be higher in countries with a pluricentric language [Clyne, 1991] and a public conscience of the many different spoken dialects. This is the case in Austria, where many speakers are competent in both, a dialect and the Austrian German standard [Moosmüller, 1997].

Our methods can also be applied to the modeling and interpolation of second language learning accents. There has been an increasing amount of work in cross-lingual speaker transformation in the last years [Wu et al., 2009a; Liang et al., 2010; Liang and Dines, 2011; Qian et al., 2011; Oura et al., 2012], which will continue to be important due to the possible applications in machine translation, personalization of speech output systems, and language learning. We have started initial investigations into the field of second language learning accent transformation [Toman and Pucher, 2015a]. These methods allow us to transform an accented speakers voice into a standard language voice such that the speaker can hear his/her own voice in a different accent.

Impact of our work on modeling of fast speech

The correct modeling of speech duration on a phonetic level has been an important task in speech synthesis since its beginning [Flanagan et al., 1970]. In our work [Pucher et al., 2010a; Valentini-Botinhao et al., 2014, 2015] we showed how to achieve a flexible modeling of fast speech with trained duration models, and evaluated the performance of blind and non-blind listeners in the decoding of fast and ultra-fast speech.

After the publication of our papers there has been continued work on the modeling of speaking rate in speech synthesis referencing our work. Hsieh et al. [2012] proposed an approach to model Mandarin-speech prosody by taking speaking rate as a continuous independent variable and letting prosodic-acoustic features depend on it. An extension of the model was proposed in Chen et al. [2014] where the model used "12 sub-models to describe various relationships of prosodic-acoustic features of the speech signal, linguistic features of the associated text, and prosodic tags representing the prosodic structure of speech".

Impact of our work on perception of synthetic speech

Studies into the perception of synthetic speech are an important knowledge source for the correct evaluation of speech synthesis systems [Pisoni, 1997]. But speech synthesis systems also provide new opportunities for investigations in speech perception, which we have also shown in our study [Pucher et al., 2015] where we investigated how listeners perceived their own or known speakers synthesized voices. We were able to show that blind and visually impaired listeners had an increased engagement time and performance when hearing their own voices. We have not seen yet much impact of this very recent work, but it will surely be interesting to the community. Meanwhile, we have also extended this first study by synthetic speaker and speech recognition experiments where we could show that blind listeners outperform their visually impaired companions in synthetic speaker recognition [Pucher et al., 2016b].

Artificial Intelligence (AI) and context modeling

With our adaptive methods it is now possible to deploy very flexible speech output systems. However, with the growing number of control parameters that these system can have now, there is also a need to know how we can steer the control parameters from a textual input. Mastering such complex control parameter spaces and thereby modeling a large amount of different contexts is a general problem in *Artificial Intelligence* (AI) [Serafini and Bouquet, 2004]. In this respect our work also shows how we can extend the control parameter space for speech synthesis through contextual modeling.

Realism in speech synthesis

The investigations of dialect or in general language variety synthesis (dialect, sociolects, accent) aims to realize realism in speech output technology. Speech interfaces that are only exhibiting one standard language are unrealistic, if we acknowledge that most of the spoken language that is uttered day-by-day is non-standard language. It may be questionable if it is desirable to achieve realistic speech output interfaces because we may arrive at an uncanny valley [Mori et al., 2012], which may need more and more effort for achieving small improvements towards realism.

Another possible criticism of realistic speech output is that we might always wish to have a clear perception of the difference between our "relational artifacts" [Turkle, 2007] and other human beings. Based on these arguments it could be argued that it is sufficient to have one good synthesizer for every standard language variety. A techno-critical argument against realistic or authentic "relational artifacts" is that they might change our general attitudes towards humans or relationships in general, because the communication with them does not lead to frustrating experiences [Turkle, 2007].

The wish to realize many different voices of a language or custom voices for specific customers [ORF, 2015] is however evidence that there is a trend towards more realism in speech output interfaces. For some special application scenarios like gaming and service robots, synthesizers that can realize multiple language varieties are also crucial. From existing trends in *Information and Communication Technology* (ICT) we can predict that an increasing level of realism will continue to be a desideratum for future speech output interfaces. Based on the achieved quality of our synthesizers these "relational artifacts" will sometimes be taken for human beings, and will sometimes leave us in an uncanny valley.

2.1.2 Audio-Visual Text-to-Speech Synthesis

Realistic 3D audio-visual text-to-speech synthesis that can be used in computer games is still an unsolved problem, although we can produce a sufficient quality to adopt the technology in real-life applications [Mattheyses and Verhelst, 2015]. Today's technology can capture high-dimensional meshes for representing faces and facial movements and can apply the captured data to 3D avatars [Facewaretech, 2015], the problem of controlling these high dimensional meshes with a model that can be learned on a low-dimensional control parameter space [Deng and Neumann, 2008], as well as the problem of synthesizing visual speech for many different contexts are still not solved [Mattheyses and Verhelst, 2015]. For solving the audio-visual text-to-speech synthesis problem we need competence from different fields such as speech communication and computer graphics [Berger et al., 2011]. This interdisciplinary nature of the problem poses an additional difficulty.

In our work on audio-visual text-to-speech synthesis we propose two methods for increasing the quality of the model, namely adaptive modeling [Schabus et al., 2012] and joint audio-visual modeling [Schabus et al., 2014], as well as one control model that allows us to control the acoustic output through visual parameters represented in a low dimensional *Principal Component Analysis* (PCA) space [Hollenstein et al., 2013].

Impact of our work on audio-visual speech synthesis

The recognition of our work in the field of visual speech synthesis is also shown by our organization of the conference "FAAVSP 2015 - The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing" [Pucher et al., 2015a,b; Pucher, 2015d] and the conference "FAA 2012 - The ACM 3rd International Symposium on Facial Analysis and Animation" in Vienna [Pucher et al., 2012a; Pucher, 2012b].

At the two days of FAAVSP 2015 we had 8 oral sessions on "Facial Analysis and Synthesis", "Perception, Emotion, and Corpora", "Life Span", "Emotion, Personality, and Dialogue", "Culture and Language", "Visual Speech Synthesis", "Audio-Visual Speech Recognition", "Visual Speech Perception", and 3 poster sessions. We also had 5 top keynote speakers (Volker Helzle Filmakademie Baden-Württemberg, Institute of Animation; Veronica Orvalho, University of Porto; Jean-Luc Schwartz, GIPSA Lab; Frank Soong & Lijuan Wang, Microsoft Research Asia).

At the 1-day event FAA2015 we also had 3 top keynote speakers (Jörn Ostermann, Leibniz Universität Hannover, Laboratory for Information Technology; Mark Pauly, Ėcole Polytechnique Fėdėrale de Lausanne (EPFL), Switzerland; Thabo Beeler, Disney Research Zurich), 7 oral presentation and 11 posters. Image Metrics and Dimensional Imaging also showed industry demos.

Our work on HMM-based audio-visual speech synthesis was referenced in several research papers on visual synthesis. Mattheyses and Verhelst [2015] gives an overview of the state-of-the-art in audio-visual speech synthesis. They classify our approach in Schabus et al. [2014] as a single-phase system that models the acoustic and visual modality jointly. According to [Mattheyses and Verhelst, 2015] such systems are the only ones that "can be considered as truly audiovisual synthesizers". Zhu et al. [2015] introduces a method based on a shift-invariant learned dictionary, which can capture the bimodal structure of articulation. In Kim et al. [2015] a sequence prediction method is discussed that predicts sequences that lie within a high-dimensional continuous output space, which is what we also have to do in speech synthesis. They use a decision tree framework for learning a non-parametric spatiotemporal sequence predictor.

Cakmak et al. [2014b] applied speaker-dependent training of Hidden Markov Models (HMMs) to audio and visual laughter synthesis. They used separate models for audio and visual signals and also showed that extrapolation of visual laughter synthesis can be done. Cakmak et al. [2015] proposes the use of synchronization rules between acoustic and visual laughter synthesis systems. This is necessary since their acoustic and visual

models are trained independently without any synchronization constraints. In Cakmak et al. [2014a] the synchronous corpus of audio and 3D facial marker data for audio-visual laughter synthesis is introduced.

2.1.3 Speaker Verification Spoofing

In our work on speaker verification spoofing [De Leon et al., 2012][De Leon et al., 2010b, 2011, 2010a] we showed how an adaptive speech synthesizer can spoof a speaker verification system. The problem of imposture against *Speaker Verification* (SV) systems using speech synthesized from HMMs was first identified in the late 90s by Masuko et al. [1999]. In their original work, the authors used an HMM-based text-prompted SV system and an HMM-based speech synthesizer. In the SV system, feature vectors were scored against speaker and background models composed of concatenated phoneme models (not GMM-based models).

When we started to work on the problem, HMM-based synthesis had made big progress compared to the late 90s such that it was possible to adapt a speech synthesizer from a small amount of data. This increased the threat for speaker verification, and shows the general structure of any synthetic data impostor problem. Whenever the verification (classification) technology and the associated synthesis (regression) technology are at a similar technological level in terms of performance, needed data, training time, and so on, a synthetic signal impostor problem can arise.

Impact of our work on speaker verification spoofing

Our first paper on the topic of *Speaker Verification Spoofing* (SVS) was published in 2010. Table 2.1 shows the number of new publications per year from 2007 until 2016 that contain the string "speaker verification spoofing" (Column 1-2) or all three strings "speaker" and "verification" and "spoofing" (Column 3-4). The concept of "speaker verification spoofing" that we did not use in our initial publications did not appear until 2011 and the two papers in 2011 and 2012 that used the concept referenced our papers from 2010. After 2012 there was an exponential increase, which shows the same trend for 2016 with already 12 publications on the topic. From this we can see that our papers were important references that initiated the field.

When we look at column 3-4 we can see that the general topic of SVS was of interest throughout the years, most papers referencing replay attacks and the like. But also from 2010 on we can see a significant awareness concerning the general topic, which is possibly also influenced from the investigations that we have made concerning spoofing by adaptive synthesizers. An exact proof of such dependencies is difficult, but we can definitely conclude that our papers were very influential to the field.

"speaker verification spoofing"		"speaker" & "verification" & "spoofing"	
Year	No. of papers	Year	No. of papers
2007	0	2007	307
2007	0	2007	304
2008	0	2008	309
2009	0	2009	392
2010	0	2010	445
2011	1	2011	494
2012	1	2012	554
2013	3	2013	718
2014	14	2014	792
2015	37	2015	948
2016	12	2016	196

Table 2.1: *Number of new papers found in Google Scholar for the query "speaker verification spoofing" (Column 1-2) and "speaker" and "verification" and "spoofing" (Column 3-4) in a certain year.*

Our papers received much attention in the speaker verification community and led to a multi-disciplinary research topic of 'voice anti-spoofing". Since the publication of our papers several initiatives have investigated the topic [Evans et al., 2013a; Marcel, 2014], with the latest initiative being the spoofing challenge "ASVspoof 2015: Automatic Speaker Verification Spoofing and Countermeasures Challenge" at Interspeech 2015 [Kinnunen et al., 2015].

The influence of our papers may also be shown by looking at the definition and the results of the 2015 spoofing challenge [Kinnunen et al., 2015]. The training data of the challenge consisted of spoofed utterances that were generated according to one of three voice conversion and two speech synthesis algorithms. In our work we have only evaluated spoofed utterances from a HMM based system and showed how we can detect such a system with phase-based features. In the spoofing challenge this was generalized to multiple synthesis and conversion systems and the task was to detect synthetic speech without knowledge of the synthesis system. The spoofing challenge did however only focus on the task of detecting synthetic speech, whereas we showed how to detect synthetic speech for one type of synthesis system and that synthetic speech can spoof a speaker verification system. The latter is the reason why one wants to develop a spoofing detection algorithm. Several of the algorithms that were submitted for the challenge were based on phase features [Sanchez et al., 2015; Liu et al., 2015b; Wang et al., 2015], which we have also been using in our work. For the detectors some groups used DNN [Villalba et al., 2015; Chen et al., 2015] where we used a simple *Gaussian Mixture Model* (GMM) in our detection algorithm. In total there were 12 submissions

to the challenge showing that the topic of "Automatic Speaker Verification: Spoofing and Countermeasures" received much attention in 2015 and has the potential to still be a challenge in the forthcoming years.

The work done before the 2015 spoofing challenge that also referenced our work on speaker verification spoofing took several directions. Wu et al. [2015a] provides a survey of spoofing and countermeasures in speaker verification. Wu et al. [2015b] introduces a speaker verification spoofing and anti-spoofing database, which contains nine spoofing techniques, two speech synthesis based, and seven voice conversion based methods. Sizov et al. [2015] focuses on voice conversion spoofing and models speaker verification and anti-spoofing jointly in the i-vector space. This enables the integration of speaker verification and anti-spoofing tasks into one system.

Sanchez et al. [2014] performed a cross-vocoder study to evaluate the performance of a synthetic speech classifier based on relative phase shift features. McClanahan et al. [2014] investigated the vulnerability of a state-of-the-art i-vector SV system against HMM-based synthetic speech from an adapted model. They showed that i-vector based systems are vulnerable to synthetic speech, as are GMM-based and *Support Vector Machine* (SVM) based systems.

Wu et al. [2013b] proposes a fusion of phase modulation features and phase features for the detection of synthetic speech. In our original paper [De Leon et al., 2012] we have also used phase features based on the harmonic model. Evans et al. [2013b] discusses different spoofing attacks and countermeasures in speaker verification. Possible spoofing methods are impersonation, replay, speech synthesis, and voice conversion, which all need different countermeasures. Wu et al. [2013a] evaluates the vulnerability of text-independent and text-dependent speaker verification systems in a single study. De Leon and Stewart [2013] proposes a method that uses specific words, which provide strong discrimination between human and synthetic speech in a text-dependent speaker verification scenario.

De Leon et al. [2012] uses pitch patterns such as mean pitch stability, mean pitch stability range, and jitter as features extracted after image analysis of pitch patterns, since they have observed that for synthetic speech, these features lie in a small and distinct space as compared to human speech.

2.2 Text-to-Speech Synthesis

2.2.1 Supervised variety interpolation

Statistical parametric speech synthesis based on hidden Markov models (HMMs) has become established and well-studied, and has an ability to generate natural-sounding synthetic speech [Yoshimura et al., 1999a; Black et al., 2007; Zen et al., 2009b]. In recent years, the HMM-based speech synthesis systems have reached performance levels comparable to state-of-the-art unit selection systems [Fraser and King, 2007; Karaiskos et al.,

2008]. In this method, acoustic features such as the spectrum, excitation parameters, and segment duration are modeled and generated simultaneously within a unified HMM framework. A significant advantage of this model-based parametric approach is that speech synthesis is far more flexible compared to conventional unit-selection methods, since many model adaptation and model interpolation methods can be used to control the model parameters and thus the characteristics of the generated speech [Yoshimura et al., 2000a; Yamagishi et al., 2009b]. In fact, these methods have already been applied to generating transitions between different speakers [Yoshimura et al., 2000a], different types of emotional speech, and different speaking styles [Tachibana et al., 2005b].

These techniques are also useful for achieving *varying* multi-dialect voices in text-to-speech (TTS) synthesis. They may be used for personalizing speech synthesis systems and have several potential benefits. For example, if the TTS system is used to provide an alternative voice output for patients who have progressive dysarthria [Creer et al., 2009], some patients will desire a TTS system that has the same dialect as themselves.

However, it is not always feasible to prepare pronunciation dictionaries separately for every possible language variety in advance, since writing dictionaries is an extremely time-consuming and costly process. Often one variety is taken as a standard, and the linguistic resources such as pronunciation dictionaries are only available for this standard variety. Thus, to flexibly model as many varieties as possible, some acoustic and linguistic control based on this standard or typical dialect is required.

Although one might regard dialect interpolation[1] as conceptually equivalent to emotional interpolation mentioned above, there is a significant difference in the requirements for the control of dialectal varieties. The speaker or emotional interpolation mentioned above implicitly assumes that the target models use the same pronunciation dictionary, and therefore phone strings, within the same language and linear interpolation is applied just to the relevant models, which results in acoustic transitions within the same phone or sub-word unit. For dialect control, we need to additionally consider linguistically-motivated transitions. In other words, we need to include not only the HMMs but also the pronunciation dictionary as targets of the interpolation process. That is, the HMMs to be interpolated may represent different phone sequences derived from different dictionaries. Moreover, these sequences may also consist of a different number of phones.

A major premise for dialect interpolation is that dialects, as varieties of languages, form a "continuum" [Saussure, 1983]: the varieties are related to one another in terms of being linguistically close, which makes it possible for us to hypothesize the existence of varieties on that continuum of fine-grained subtleties that lie between two different varieties already defined by linguistic resources. In addition to geographical transition

[1] In our work on supervised interpolation we use the notion of 'dialect' in a broad sense as referring to non-standard language varieties. In the case at hand, it would be more accurate to speak of Viennese sociolect, since language varieties in Vienna are discerned by social criteria and not (or no longer) identified by association to a certain geographical region. We use the term 'dialect interpolation' as shorthand for 'interpolation of dialectal or sociolectal language variety'.

of the dialect varieties, that is, regiolects, we may apply the same logic to other varieties of languages such as sociolects, which are categories of linguistic varieties defined by the social level of speakers.

The proposed dialect interpolation aims to produce synthetic speech in a phonetically intermediate variety from given models and dictionaries for adjacent typical varieties. For the phonetic control, we simply use linear interpolation of HMMs that represent the acoustic features similar to speaker or emotional interpolation. Since relations between articulatory and acoustic features are non-linear [Stevens, 1997], the phonetic control that can be achieved using acoustic features alone is noisy and might sometimes exhibit unexpected behavior. However it is worthwhile to investigate the basic performance of acoustic interpolation because proper acquisition of articulatory features requires specialized recording equipment such as electromagnetic articulography (EMA) [Schönle et al., 1987] and also because phonetic knowledge such as vowel height or backness and place or manner of articulation can be used in clustering the acoustic HMMs via manually-defined linguistic questions.

A closer inspection of potential phonetic transitions between language varieties reveals several exceptional cases. From phonetic studies of Viennese dialects [Moosmüller, 1987] we know that some gradual transitions are well motivated (e.g., spirantization of intervocalic lenis plosives), while some other transitions between phones are strong markers for that specific variety, and thereby categorical. In the latter case, either the standard form of a given phone is produced, or its dialectal counterpart, with no possible in-between variants. One example of such a transition is the phone [aː] in the Standard Austrian German variety which is realized as [ɔ̝ː] in the Viennese dialect. For such a case, the use of interpolation (e.g., model interpolation between [aː] and [ɔ̝ː] phone HMMs) is not appropriate. For this reason, we introduce several knowledge-based switching rules that allow for overriding acoustic interpolation in such cases. Since it is known from psycholinguistics that continuous transitions between phones are often only perceived categorically [Liberman, 1970], the knowledge-based switching rules should improve the perception of dialects compared to acoustic interpolation alone. Hence, we include interpolations with and without switching rules in the subjective evaluation to measure the effect of the proposed dialect interpolation and switching rules.

In addition we investigate efficient clustering strategies for the dialect varieties in HMM-based speech synthesis. In general there are insufficient speech resources for non-standard dialect varieties. This situation might be even more severe for minor languages. Thus we compare several clustering algorithms for a practical case where the amount of speech data for dialects is limited, but there is sufficient speech data for the standard. We also include speech data from speakers that are able to speak standard and dialect.

2.2.2 Unsupervised variety interpolation

The flexibility of Hidden Semi-Markov Model (HSMM) based speech synthesis allows for different strategies to manipulate the trained models, such as adaptation and interpolation. We develop, analyze, and evaluate unsupervised interpolation methods that can be used to generate intermediate stages of two language varieties. "Variety" is a cover term void of any positive or negative evaluative connotations. It comprises dialects, sociolects, and standard languages. In this contribution, we apply this method to perform an interpolation between Regional Standard Austrian German (RSAG) and three dialects/sociolects. The difficulty of dialect interpolation lies in lexical, phonological, and phonetic differences between the varieties [Russell et al., 2013]. In this contribution we focus on interpolation of phonetic differences.

In recent years there have been several research efforts in the context of language varieties for speech synthesis, reviewed in Russell et al. [2013]. Following Russell et al. [2013] we can distinguish between fully-resourced and under-resourced modeling as well as different applications like variety interpolation.

In fully-resourced modeling, Richmond et al. [2010] described how to generate pronunciation dictionaries based on morphological derivations of known words. They reported that in preliminary experiments for 75% of tested words, their method produced the correct, fully-specified transcription. This can be used as an extension to existing grapheme-to-phoneme rules to obtain contextual information on out-of-vocabulary words and could be beneficial for building an actual dialect synthesis system that includes interpolation.

Nguyen et al. [2013] described the development of an HMM-based synthesizer for the modern Hanoi dialect of Northern Vietnamese, describing special challenges they encountered, comparable to our process of acquiring our dialect corpus.

In Toman et al. [2013c] we evaluated different acoustic modeling methods for dialect synthesis. The interpolation technique presented in the present work is compatible to all acoustic modeling methods as long as they produce a HSMM state sequence for a given set of labels.

For developing synthesizers for under-resourced languages, different methods have been developed to aid the process of data acquisition and annotation.

Goel et al. [2010] evaluated the combination of different lexicon learning techniques with a smaller lexicon available for bootstrapping. In their experiments, their method could increase the Word Recognition Accuracy from 41.38% for a small bootstrap lexicon to 43.25%, compared to 44.35% when using the full training dictionary.

Watts et al. [2013] developed methods and tools for (semi-)automatic data selection and front-end construction for different languages, varieties and speaking styles e.g. from audio books. Results from Watts et al. [2013] are published by Stan et al. [2013] who applied these tools on "found speech" to create a standardized multilingual corpus.

For our work on dialectal synthesis, such methods are useful for easy acquisition and annotation of dialect data, which is currently a time-consuming process.

Loots and Niesler [2011] developed a phoneme-to-phoneme conversion technique that uses decision trees to automatically convert pronunciations between American, British and South African English accents[2]. This method could be used to automatically generate the phonetic transcription for less-resourced dialects from a fully-resourced variety, as a transcription of the dialect utterance is required for our interpolation technique presented here.

Voice model interpolation was first applied in HSMM-based synthesis for speaker interpolation [Yoshimura et al., 2000b] and emotional speech synthesis [Tachibana et al., 2005a]. Picart et al. [2011] used model interpolation to create speech with different levels of articulation. Lecumberri et al. [2014] considered the possibility of using extrapolation techniques to emphasize foreign accent as an application for foreign language learning. The methods presented here could also be used to produce an extrapolated dialect, but this is not investigated in our work.

In language variety interpolation, Astrinaki et al. [2013] have shown how to interpolate between clusters of accented English speech within a reactive HMM-based synthesis system. In this method, phonetic differences between the accent representations were not considered (i.e. the same set of phone symbols and utterance transcriptions was used for all accents).

In the previous section we have described how to interpolate between phonetically different dialects in a supervised way. In this method, we used a manually defined phone-mapping between Standard Austrian German and the Viennese dialect. Evaluation tests showed that listeners actually perceive the intermediate varieties created by interpolation as such.

In this contribution we extend the method from Pucher et al. [2010b] to work in an unsupervised way, such that no manually defined mapping is necessary, therefore allowing the fully automatic interpolation. Also, interpolation is performed between RSAG and three dialects/sociolects. This unsupervised method is based on Dynamic Time Warping (DTW) [Rabiner et al., 1978] on HSMM state level. Compared to Pucher et al. [2010b], this method introduces one-to-many mappings between states, requiring a more sophisticated duration modeling procedure.

To introduce the integration of phonological knowledge in the interpolation technique, we describe the following alternations, which characterize the RSAG - dialect interaction[3]:

1. **Phonological process**: Socio-phonological studies on Austrian varieties demonstrate that certain alternations between two varieties, usually a standard variety

[2]The term "accent" is often used for regional differences of English. We avoid the term "accent" in this contribution as it refers to more than one linguistic phenomenon and we specifically treat dialects here.

[3]// denotes the phonological representation, [] the phonetic realization.

and a dialect, are phonetically well motivated and thus can be described as phonological processes, e.g., spirantization of intervocalic lenis stops [Moosmüller, 1991] like

- [ɑːb̥ɐ] to [ɑːbɐ] to [ɑːβɐ]
  aber (engl. "but") or
- [laɛd̥ɐ] to [laɛdɐ] to [laɛðɐ]
  leider (engl. "unfortunately").

Interpolation can be used to model these gradual transitions.

2. **Input-switch rules**: Other alternations lack such phonetic motivations because of a different historical development. These alternations are therefore described as input-switch rules, e.g.
   - /g̊uːt/ ↔ /guɑd̥/ or
   - /g̊uːt/ ↔ /g̊uid̥/
     gut (engl. "good").

   No gradual transitions from e.g., /g̊uːt/ to /guɑd̥/ can be observed [Dressler and Wodak, 1982; Moosmüller, 1991]. Because of their phonetic saliency, input-switch rules are sociolinguistic markers as defined by Labov [1972], meaning they are subjected to stereotyping and social evaluation (positive or negative). Therefore, interpolation is not feasible in these cases.

3. **Pseudo-phonological process**: Many input-switch rules involve diphthongs vs. monophthongs; i.e. the standard form is a diphthong, the dialect form is a monophthong. Standard Austrian German features a vast variety of phonetic diphthongal realizations (Moosmüller *et al.*, in press), so that any (slight) movement in formant frequencies is interpreted as a diphthong [Moosmüller and Vollmann, 2001]. Sociolinguistically, the input-switch rule persists; the diphtong is the standard form, the monophthong is the dialect form, such as in the following examples:
   - /haɛs/ ↔ /hɑːs/
     heiß (engl. "hot") or
   - /kɑɔ̯fs̊d̥/ ↔ /kɑːfs̊d̥/
     kaufst (engl. "(you) buy")

   However, the gradual decrease in formant frequency movement can be elegantly captured by interpolation, without attracting negative evaluation from the listener's part. Consequently, modeling this case using HSMM interpolation is feasible although the alternation is actually an input-switch rule.

When input-switch rules are considered, it is not phonetically feasible to interpolate whole utterances. Therefore, we introduce region-based interpolation. This introduces another level of mappings on regions spanning multiple phones. These regions can then be defined as either (pseudo-)phonological process or input-switch rule. For example, the words **Ziege** (RSAG) vs. **Goaß** (dialect; engl. "goat") might form mapped regions that should not be interpolated.

The developed interpolation methods have possible applications in spoken dialog systems where we aim to adapt speech output to the user of the dialog system. As soon as the dialect/sociolect of the user is detected, we can use interpolation to create a dialog system persona that fits the dialect/sociolect spoken by the user. In Toman and Pucher [2013]; Toman et al. [2013b], we presented a method for cross-variety speaker transformation based on HSMM state mapping [Wu et al., 2009b]. Transforming the voice of a speaker from one variety to another can be used as a basis for dialect interpolation. For example, a single voice model could be transformed to multiple other varieties and then interpolation can be used to synthesize samples for intermediate stages, enabling a large spectrum of speaking styles. Furthermore, interpolation methods could also be used to extend existing multi-variety speech databases or speech databases with similar languages by augmenting them with interpolated data. In general, our methods can be applied to any interpolation of state sequences of HMM models, which makes it also applicable for facial animation [Schabus et al., 2014].

Our HSMM-based synthesizer is an extension of the HSMM-based speech synthesis system published by the EMIME project [Yamagishi and Watts, 2010]. The methods for training these kinds of synthesizers and synthesizing from HSMMs were published in a number of papers [Zen et al., 2009b; Tokuda et al., 1995; Yoshimura et al., 1999b; Yamagishi and Kobayashi, 2007a; Tokuda et al., 1999].

2.2.3 Modeling of fast speech

It is well known that synthetic speech at very high speaking rates is frequently used by blind users to increase the amount of presented information. In data-driven approaches, however, this may lead to a severe degradation of synthetic speech quality, especially at very fast speaking rates. The standard HSMM-based duration modeling Zen et al. [2007b] is already able to model certain non-linearities between normal and fast speech units since it uses explicit state duration distributions and can thereby take the duration variance of units into account. But for very fast speaking rates this is not sufficient. We therefore propose a duration control method using a model interpolation technique, where we can continuously interpolate HSMMs for normal and fast speaking rate. The HSMMs for fast speaking rate are adapted from HSMMs for normal speaking rate. In addition to interpolation between normal and fast speaking rate, we can also use extrapolation between models to achieve very fast speaking rates that go beyond the recorded original speaking rates. A conventional study Iwano et al. [2002] already showed that an HMM-based synthesizer with interpolated duration models can outperform a

synthesizer with rule based duration model. Their models were, however, based on the so called Hayashi's quantification method I and were theoretically different from our methods that are based on HSMM interpolation and adaptation techniques, which are available from the HTS toolkit today Zen et al. [2009a].

Some studies have shown that the complex duration changes between normal and fast speech are present at several linguistic levels Janse [2004]. Therefore we employ context-dependent linear regression functions for the HSMM duration adaptation to model the duration changes at different linguistic levels. The contexts we used also include high-level linguistic features such as syllable information, phrase information etc. The use of HSMM duration adaptation has another advantage. It makes online processing of the proposed duration control technique possible since normal and fast duration models have the same tying structure and we can straightforwardly perform the interpolation online. This also makes the analysis of the modeling error of standard duration modeling for fast durations easier.

For the evaluation we carried out a comprehension and pair wise comparison test with both blind and sighted listeners. We confirmed that both groups of listeners preferred sentences generated with our method than the conventional method. The proposed method could also achieve lower word error rates (WER) in the comprehension test. The blind listeners were especially good in understanding sentences at fast speaking rates (8-9 syllables per second) compared to non-blind listeners.

Blind individuals are capable of understanding speech reproduced at considerably high speaking rates Moos and Trouvain [2007]. As screen readers become an essential computer interface for blind users, a challenge arises: how to provide intelligible synthesized speech at such high rates? The standard HSMM-based synthesizer Zen et al. [2007b] models speech duration by using explicit state duration distributions but for very fast speaking rates this is often not sufficient Pucher et al. [2010a]. It is also unclear whether using fast speech to train a synthesizer can create more intelligible fast synthesized speech than other sorts of compression methods.

Fast speech production and perception has been the target of various studies Gay [1978]; Janse et al. [2003]; Port [1981]; Goldman-Eisler [1968]; Greisbach [1992]. When producing fast speech vowels are compressed more than consonants Gay [1978] and both word-level Janse et al. [2003] and sentence-level Port [1981] stressed syllables are compressed less than unstressed ones. Yet another important aspect of fast speech is the significant reduction of pauses. It is claimed that reducing pauses is in fact the strongest acoustic change when speaking faster Goldman-Eisler [1968], most probably due to the limitations of how much speakers can speed up their articulation rate Greisbach [1992]. It is argued that these observed changes are the result of an attempt to preserve the aspects of speech that carry more information. The presence of pauses however have been shown to contribute to intelligibility Sanderman and Collier [1997].

It has been shown that fast speech (around 1.56 times faster than normal speech) is harder to process, in terms of reaction time, and also preferred less than linearly com-

pressed speech Janse et al. [2003]; Janse [2004]. Linearly compressed speech was found to be more intelligible and better liked than a nonlinearly compressed version of speech where fast speech prosodic patterns were mimicked Janse et al. [2003]. The author claims that possibly the only nonlinear aspect of natural fast speech duration changes that can improve intelligibility at high speaking rates is pause removal but only when rates are relatively high Janse [2004]. Another nonlinear compression method is the MACH1 algorithm Covell et al. [1998]. This method is also based on the acoustics of fast speech with the addition of compressed pauses. It has been shown that at high speaking rates (2.5 and 4.1) MACH1 improves comprehension and is preferable to linearly compressed speech but no advantage was found at the fast speech speaking rate (1.4) He and Gupta [2001].

Fast synthesized speech generated by a formant-based system was found to be less intelligible than fast natural speech and the intelligibility gap grows with the speaking rate Lebeter and Saunders [2010]. More recently the authors in Syrdal et al. [2012] evaluated the intelligibility of a wider range of synthesizers: formant, diphone, unit selection and HMM-based. It was found that the unit selection systems were more intelligible across speech rates. In this evaluation, however, the evaluated synthesizers were based on different speakers and the compression methods adopted by each system were not reported. Literature on fast synthesized speech also focuses on the effect on blind listeners. To improve duration control of HMM-based systems for blind individuals Pucher et al. [2010a] proposed a model interpolation method. Pucher et al. found that interpolating between a model trained with normal and a model trained with fast speech data results in speech that is more intelligible and preferable, for both blind and non blind individuals.

In our work, we are interested in analysing two aspects of fast synthesized speech. First, the corpus used to train synthesis models, i.e., is it really necessary or even helpful to use fast speech recordings? Second, compression method; which is more effective: a nonlinear manipulation of speech duration or a linear compression method? We evaluate intelligibility of a fast and a normal female Scottish voice and a German male voice, compressed using two nonlinear and one linear method and presented to listeners at different rates.

2.2.4 Perception of synthetic speech

There is an ever increasing amount of applications that require customised speech synthesis that can reflect accent, speaking style and other features, particularly in the area of assistive technology [Pucher et al., 2010b][Yamagishi et al., 2012]. Current speech technology techniques make it possible to create synthetic voices that sound considerably similar to the original speaker using only a limited amount of training data Yamagishi and Kobayashi [2007b]. This naturally leads to research questions regarding how a listener's perception of a synthetic voice depends on the listener's acquaintance with the speaker used to train the voice. Moreover how does one perceive a synthetic voice trained on one's own speech. These questions are particularly of interest when considering the

design of audio lecture material for blind children and how learning may be improved by using familiar voices. One idea we are looking to exploit is the impact of using the child's own voice or that of their teacher.

To the best of our knowledge there are no existing studies on the perception of one's own synthetic voice. Studies on the perception of one's own natural voice exist but are quite sparse and do not report on preference or intelligibility results Fernyhough and Russell [1997]; Appel and Beerends [2002]; Rosa et al. [2008]. There is however an extensive literature on the perception of familiar voices Van Lancker et al. [1985]; Lancker and Kreiman [1987]; Bőhm and Shattuck-Hufnagel [2007]; Nygaard et al. [1994]; Nygaard and Pisoni [1998]; Yonan and Sommers [2000]; Newman and Evers [2007]; Souza et al. [2013]. Most studies create familiarity by exposing their listeners to a certain voice, either in one or a few sessions across a certain time range Nygaard et al. [1994]; Nygaard and Pisoni [1998]; Yonan and Sommers [2000]. Such studies found that for both young adults Nygaard et al. [1994]; Nygaard and Pisoni [1998] and older adults Yonan and Sommers [2000] prior exposure to a talker's voice facilitates understanding. In fact it's argued that this facilitation occurs because familiarity eases the effort for speaker normalization, i.e. the mapping of an acoustic realization produced by a certain speaker to a phonetic representation Pisoni and Remez [2008]. Relatively few studies evaluated the impact of long-term familiarity, i.e., a voice you have been exposed to for weeks, months or years Newman and Evers [2007]; Souza et al. [2013]. Newman and Evers Newman and Evers [2007] report an experiment of pupils shadowing a teacher's voice in the presence of a competing talker. Results show that pupils that were made aware that the target voice was their teacher's outperformed pupils that were unaware of this or that were unfamiliar with that particular teacher. Souza and colleagues Souza et al. [2013] measured the long-term familiarity impact on speech perception by selecting spouses or pairs of friends and measuring how well they understand each other in noise. They found that speech perception was better when the talker was familiar regardless of whether the listeners were consciously aware of it or not.

There are also studies on the effect of familiarity of synthetic voices using a variety of synthesisers Reynolds et al. [2000b]. It has been shown that increased exposure to synthetic speech improves its process in terms of reaction time Reynolds et al. [2000b]. There are far fewer studies on the perception of synthetic speech which is similar to a particular person's voice or that has been synthesized with a particular voice Nass and Lee [2001]; Wester and Karhila [2011]. Nass and Lee [2001] showed that synthetic voices that are acoustically similar to one's own voice are generally not preferred over non-similar voices. A preference was however found for voices that showed the same personality as defined by duration, frequency, frequency range, and loudness of the voice. Another study Wester and Karhila [2011] showed that it is more difficult for listeners to judge whether two sentences are spoken by the same person if one of the sentences is produced by a speech synthesizer and the other is natural speech as opposed to both being synthetic speech.

It has been shown that blind individuals obtain higher intelligibility scores when com-

pared to sighted individuals Hugdahl et al. [2004] and that this benefit is also observed for the intelligibility of synthetic speech Papadopoulos et al. [2008] Pucher et al. [2010a] possibly due to the familiarity effect Barouti et al. [2013] as blind individuals are exposed to the material more through the use of screen readers and audio books.

In the context of a research project together with a school for blind children we evaluated the use of different synthetic voices in audio games. Assuming that synthetic voices still benefit from the familiarity effect and that one's own synthetic voice is in a certain way a familiar voice, we evaluate the engagement time and game performance of a group of blind children playing audio games incorporating their own synthetic voice, their teacher's synthetic voice and an unknown synthetic voice. Using a HMM-based speech synthesis system for German we built voices of 18 school children and 7 teachers of the same school and an additional speaker who was not known to the children.

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 72 (2015) 176–193

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Unsupervised and phonologically controlled interpolation of Austrian German language varieties for speech synthesis

Markus Toman [a,*], Michael Pucher [a], Sylvia Moosmüller [b], Dietmar Schabus [a]

[a] Telecommunications Research Center Vienna (FTW), Donau-City-Str 1, 3rd floor, 1220 Vienna, Austria
[b] Austrian Academy of Sciences – Acoustics Research Institute (ARI), Wohllebengasse 12-14, 1st Floor, 1040 Vienna, Austria

Received 10 December 2014; received in revised form 18 May 2015; accepted 4 June 2015
Available online 12 June 2015

Abstract

This paper presents an unsupervised method that allows for gradual interpolation between language varieties in statistical parametric speech synthesis using Hidden Semi-Markov Models (HSMMs). We apply dynamic time warping using Kullback–Leibler divergence on two sequences of HSMM states to find adequate interpolation partners. The method operates on state sequences with explicit durations and also on expanded state sequences where each state corresponds to one feature frame. In an intelligibility and dialect rating subjective evaluation of synthesized test sentences, we show that our method can generate intermediate varieties for three Austrian dialects (Viennese, Innervillgraten, Bad Goisern). We also provide an extensive phonetic analysis of the interpolated samples. The analysis includes input-switch rules, which cover historically different phonological developments of the dialects versus the standard language; and phonological processes, which are phonetically motivated, gradual, and common to all varieties. We present an extended method which linearly interpolates phonological processes but uses a step function for input-switch rules. Our evaluation shows that the integration of this kind of phonological knowledge improves dialect authenticity judgment of the synthesized speech, as performed by dialect speakers. Since gradual transitions between varieties are an existing phenomenon, we can use our methods to adapt speech output systems accordingly.

Keywords: HMM-based speech synthesis; Interpolation; Austrian German; Innervillgraten dialect; Bad Goisern dialect; Viennese dialect

1. Introduction

The flexibility of Hidden Semi-Markov Model (HSMM) based speech synthesis allows for different strategies to manipulate the trained models, such as adaptation and interpolation. In this paper we develop, analyze, and evaluate unsupervised interpolation methods that can be used to generate intermediate stages of two language varieties. "Variety" is a cover term void of any positive or negative evaluative connotations. It comprises dialects, sociolects, and standard languages. In this contribution, we apply this method to perform an interpolation between Regional Standard Austrian German (RSAG) and three dialects/sociolects. The difficulty of dialect interpolation lies in lexical, phonological, and phonetic differences between the varieties (Russell et al., 2013). In this contribution we focus on interpolation of phonetic differences.

In recent years there have been several research efforts in the context of language varieties for speech synthesis, reviewed in Russell et al. (2013). Following Russell et al. (2013) we can distinguish between fully-resourced and under-resourced modeling as well as different applications like variety interpolation.

In fully-resourced modeling, Richmond et al. (2010) described how to generate pronunciation dictionaries based

* Corresponding author.
E-mail address: toman@ftw.at (M. Toman).

http://dx.doi.org/10.1016/j.specom.2015.06.005
0167-6393/

on morphological derivations of known words. They reported that in preliminary experiments for 75% of tested words, their method produced the correct, fully-specified transcription. This can be used as an extension to existing grapheme-to-phoneme rules to obtain contextual information on out-of-vocabulary words and could be beneficial for building an actual dialect synthesis system that includes interpolation.

Nguyen et al. (2013) described the development of an HMM-based synthesizer for the modern Hanoi dialect of Northern Vietnamese, describing special challenges they encountered, comparable to our process of acquiring our dialect corpus.

In Toman et al. (2013b) we evaluated different acoustic modeling methods for dialect synthesis. The interpolation technique presented in the present work is compatible to all acoustic modeling methods as long as they produce a HSMM state sequence for a given set of labels.

For developing synthesizers for under-resourced languages, different methods have been developed to aid the process of data acquisition and annotation.

Goel et al. (2010) evaluated the combination of different lexicon learning techniques with a smaller lexicon available for bootstrapping. In their experiments, their method could increase the Word Recognition Accuracy from 41.38% for a small bootstrap lexicon to 43.25%, compared to 44.35% when using the full training dictionary.

Watts et al. (2013) developed methods and tools for (semi-)automatic data selection and front-end construction for different languages, varieties and speaking styles e.g. from audio books. Results from Watts et al. (2013) are published by Stan et al. (2013) who applied these tools on "found speech" to create a standardized multilingual corpus. For our work on dialectal synthesis, such methods are useful for easy acquisition and annotation of dialect data, which is currently a time-consuming process.

Loots and Niesler (2011) developed a phoneme-to-phoneme conversion technique that uses decision trees to automatically convert pronunciations between American, British and South African English accents.[1] This method could be used to automatically generate the phonetic transcription for less-resourced dialects from a fully-resourced variety, as a transcription of the dialect utterance is required for our interpolation technique presented here.

Voice model interpolation was first applied in HSMM-based synthesis for speaker interpolation (Yoshimura et al., 2000) and emotional speech synthesis (Tachibana et al., 2005). Picart et al. (2011) used model interpolation to create speech with different levels of articulation. Lecumberri et al. (2014) considered the possibility of using extrapolation techniques to emphasize foreign accent as an application for foreign language learning. The methods presented here could also be used to produce an extrapolated dialect, but this is not investigated in the current paper.

In language variety interpolation, Astrinaki et al. (2013) have shown how to interpolate between clusters of accented English speech within a reactive HMM-based synthesis system. In this method, phonetic differences between the accent representations were not considered (i.e. the same set of phone symbols and utterance transcriptions was used for all accents).

In Pucher et al. (2010), we have shown how to interpolate between phonetically different dialects in a supervised way. In this method, we used a manually defined phone-mapping between Standard Austrian German and the Viennese dialect. Evaluation tests showed that listeners actually perceive the intermediate varieties created by interpolation as such.

In this contribution we extend the method from Pucher et al. (2010) to work in an unsupervised way, such that no manually defined mapping is necessary, therefore allowing the fully automatic interpolation. Also, interpolation is performed between RSAG and three dialects/sociolects. This unsupervised method is based on Dynamic Time Warping (DTW) (Rabiner et al., 1978) on HSMM state level and is subsequently described in Section 3. Compared to Pucher et al. (2010), this method introduces one-to-many mappings between states, requiring a more sophisticated duration modeling procedure, which will be described in Section 4.

To introduce the integration of phonological knowledge in the interpolation technique, we describe the following alternations, which characterize the RSAG – dialect interaction[2]:

1. **Phonological process**: Socio-phonological studies on Austrian varieties demonstrate that certain alternations between two varieties, usually a standard variety and a dialect, are phonetically well motivated and thus can be described as phonological processes, e.g., spirantization of intervocalic lenis stops (Moosmüller, 1991) like
   - [ɑːb̥ɐ] to [ɑːbɐ] to [ɑːβɐ]
     aber (engl. "but") or
   - [laɛd̥ɐ] to [laɛdɐ] to [laɛäɐ]
     leider (engl. "unfortunately").

   Interpolation can be used to model these gradual transitions.
2. **Input-switch rules**: Other alternations lack such phonetic motivations because of a different historical development. These alternations are therefore described as input-switch rules, e.g.

[1] The term "accent" is often used for regional differences of English. We avoid the term "accent" in this contribution as it refers to more than one linguistic phenomenon and we specifically treat dialects here.

[2] // denotes the phonological representation, [] the phonetic realization.

- /gu:t/ ↔ /guɑd̥/ or
- /g̊u:t/ ↔ /g̊u:id̥/
  g̊ut (engl. "good").

No gradual transitions from e.g., /gu:t/ to /guɑd̥/ can be observed (Dressler and Wodak, 1982; Moosmüller, 1991). Because of their phonetic saliency, input-switch rules are sociolinguistic markers as defined by Labov (1972), meaning they are subjected to stereotyping and social evaluation (positive or negative). Therefore, interpolation is not feasible in these cases.

3. **Pseudo-phonological process**: Many input-switch rules involve diphthongs vs. monophthongs; i.e. the standard form is a diphthong, the dialect form is a monophthong. Standard Austrian German features a vast variety of phonetic diphthongal realizations (Moosmüller et al., 2015), so that any (slight) movement in formant frequencies is interpreted as a diphthong (Moosmüller and Vollmann, 2001). Sociolinguistically, the input-switch rule persists; the diphthong is the standard form, the monophthong is the dialect form, such as in the following examples:

   - /haɐ̯s/ ↔ /hɑ:s/
     heiß (engl. "hot") or
   - /kɑɐ̯fsd̥/ ↔ /kɑ:fsd̥/
     kaufst (engl. "(you) buy")

   However, the gradual decrease in formant frequency movement can be elegantly captured by interpolation, without attracting negative evaluation from the listener's part. Consequently, modeling this case using HSMM interpolation is feasible although the alternation is actually an input-switch rule.

When input-switch rules are considered, it is not phonetically feasible to interpolate whole utterances. Therefore, we introduce region-based interpolation. This introduces another level of mappings on regions spanning multiple phones. These regions can then be defined as either (pseudo-)phonological process or input-switch rule. For example, the words **Ziege** (RSAG) vs. **Goa**ß (dialect; engl. "goat") might form mapped regions that should not be interpolated. This procedure is described in detail in Section 6.

The developed interpolation methods have possible applications in spoken dialog systems where we aim to adapt speech output to the user of the dialog system. As soon as the dialect/sociolect of the user is detected, we can use interpolation to create a dialog system persona that fits the dialect/sociolect spoken by the user. In Toman et al. (2013a), we presented a method for cross-variety speaker transformation based on HSMM state mapping (Wu et al., 2009). Transforming the voice of a speaker from one variety to another can be used as a basis for dialect interpolation. For example, a single voice model could be transformed to multiple other varieties and then interpolation can be used to synthesize samples for intermediate stages, enabling a large spectrum of speaking styles. Furthermore, interpolation methods could also be used to extend existing multi-variety speech databases or speech databases with similar languages by augmenting them with interpolated data. In general, our methods can be applied to any interpolation of state sequences of HMM models, which makes it also applicable for facial animation (Schabus et al., 2014).

Our HSMM-based synthesizer is an extension of the HSMM-based speech synthesis system published by the EMIME project (Yamagishi and Watts, 2010). The methods for training these kinds of synthesizers and synthesizing from HSMMs were published in a number of papers (Zen et al., 2009; Tokuda et al., 1995; Yoshimura et al., 1999; Yamagishi and Kobayashi, 2007; Tokuda et al., 1999).

This contribution is organized as follows: Section 2 describes the corpora and associated phone sets which were used in this work. Section 3 then presents the details of the interpolation methods used to generate intermediate language varieties. Duration modeling in these methods is described in Section 4. Section 5 presents a phonetic analysis of interpolated samples. Rules derived from these results are then incorporated in an extended interpolation method, described in Section 6. Section 7 describes the evaluations we conducted to assess and compare the presented methods. Finally Section 8 discusses results and concludes the work.

2. Corpora

The work presented here is based on a corpus consisting of three Austrian German dialects: the dialect of Innervillgraten (IVG) in Eastern Tyrol, the dialect of Bad Goisern (GOI) in the South of Upper Austria, and the dialect of Vienna (VD). IVG belongs to the South Bavarian dialect group, VD to the Middle Bavarian dialect group, and GOI belongs to the (South)-Middle Bavarian dialect group.

SAG refers to the variety spoken by the upper social classes of the big cultural centers located predominantly in the Middle Bavarian region (Moosmüller, 1991, 2015; Soukup and Moosmüller, 2011). Since the IVG and GOI speakers were genuine dialect speakers, meaning that they were raised in the respective dialect and learned SAG only in school, SAG spoken by these speakers contained also regional features. Therefore, the SAG variety produced by the GOI and IVG speakers is referred to as regional standard Austrian German (RSAG). In Table 1, the difference between SAG, RSAG as spoken in Bad Goisern, and GOI is illustrated.

In RSAG, /ɛ/ of **Schwester** is slightly diphthongized, a process which is not allowed in SAG. Also, the diphthong in **meine** differs between the two varieties. These two features cue a regional variant of SAG which, in turn, is still

Table 1
Differences between SAG, RSAG as spoken in Bad Goisern, and GOI.

SAG orth.	**Morgen kommt meine Schwester.**
SAG phon.	ˈmɔɐ̯gŋ̍ kɔmt maɐ̯nɛ ˈʃvɛstɐ
RSAG phon.	ˈmɔɐ̯gŋ̍ kɔmt maɐ̯nɛ ˈʃvᵉestɐ
GOI phon.	ˈmɔʀɪŋ̍ kimd̥ mãɐ̯̃ ˈʃvᵉesd̥a

completely different from the dialect version of the model sentence.

Ten dialect speakers, gender balanced, were recruited for the GOI and the IVG corpus, respectively. The recordings consisted of spontaneous speech, reading tasks, picture naming tasks, and translation tasks from SAG into the dialect. From these recordings, 660 phonetically balanced sentences were selected and a phone set was created for each dialect. For the recording of the 660 phonetically balanced dialect sentences the dialect speakers heard the dialect speech sample they were asked to utter and were also presented with an orthographic transcription of the sentence which was close to the standard language. In addition, these speakers also read a corpus of SAG sentences. The speaker selection and recording process for IVG and GOI has been described in detail in Toman et al. (2013b).

The corpus of the VD speakers is different, in as for the synthesis of the Viennese dialects and sociolects, actors and actresses were recruited. 10 actors and actresses were invited for a casting in which they had to perform reading tasks in both SAG and VD. For the VD samples, they had to transform SAG sentences into the VD. Subsequently, the recordings were subjected to analysis and the speaker who performed best was chosen for the VD dialect recording sessions. The speaker selection and recording process for VD has been described in detail in Pucher et al. (2012).

Sound samples were recorded at 44,100 Hz, 16 bits/sample. The training process was also performed using these specifications. Cutting and selection was performed manually. Noise cancellation and volume normalization was applied to the recordings. Synthesized samples used in the evaluation were also volume normalized. A 5 ms frame shift was used for the extraction of 40-dimensional mel-cepstral features, fundamental frequency and 25-dimensional band-limited aperiodicity (Kawahara et al., 1999) measures. Speaker-dependent models were trained for the evaluations using the HSMM-based speech synthesis system published by the EMIME project (Yamagishi and Watts, 2010). The interpolation methods presented in this contribution were integrated into this system.

Table 2 shows a sample of the utterances which were used for the evaluation, and which are also linguistically and phonetically analyzed in Section 5. We interpolate between Regional Standard Austrian German (RSAG) and one dialect variety (Innervillgraten dialect (IVG), Bad Goisern dialect (GOI) or Viennese dialect (VD)). In total, we use 6 different utterances per variety. There are significant phonetic and lexical differences between RSAG and the respective dialect (IVG, GOI, VD). These differences lead to different numbers of phones and

Table 2
Sample sentences which were interpolated between Regional Standard Austrian German (RSAG), Innervillgraten dialect (IVG), Bad Goisern dialect (GOI) and Viennese dialect (VD).

SAG orth.	**Schnee liegt im Garten.**
RSAG phon.	ˈʃneː liːkt ʔɪm ˈgɑːd̥n̩
IVG phon.	ˈʃneą liːd̥ iŋ̍ ˈg̥ɔχd̥ɛ
SAG orth.	**Morgen kommt meine Schwester.**
RSAG phon.	ˈmɔɐ̯gŋ̍ kɔmt maɐ̯nɛ ˈʃvᵉestɐ
GOI phon.	ˈmɔʀɪŋ̍ kimd̥ mãɐ̯̃ ˈʃvᵉesd̥a
SAG orth.	**Wir sind lustige Leute.**
RSAG phon.	viɐ̯ sɪnd̥ ˈlʊsd̥ɪgɛ ˈlɒɪtʰ
VD phon.	miɐ̯ san ˈlusd̥iɣɛ ˈlæːtʰ

different numbers of lexical items between RSAG and dialects (e.g. RSAG "d̥iː ˈtaːgɛ – **die Tage**" vs. GOI "d̥ːɔːk – **die Tage**"[3] occurred in the evaluated samples).

Table 3 shows the phone sets for the different varieties. Affricates were split into two phones in (R)SAG and VD as these were defined in a previous project. For alveolar stops, a further category had to be introduced in order to capture instances which can neither be assigned to [t] nor to [d̥]. Since consonant duration is the decisive feature in differentiating stops in Austrian dialects, we symbolize this additional category as [d̥ː] (see Moosmüller and Brandstätter (2014) for a discussion).

3. Interpolation methods

This section describes the interpolation methods used to generate intermediate language varieties.

In Pucher et al. (2010) we implemented and evaluated a *supervised interpolation* method that allows for gradual interpolation between language varieties when a phoneme mapping is given. Here we extend this method by an *unsupervised interpolation* that is based on Dynamic Time Warping (DTW) (Rabiner et al., 1978; Müller, 2007) of HSMMs. This method implements gradual interpolation between varieties without a phoneme mapping.

To obtain the DTW warping path, we use the Kullback–Leibler Divergence (KLD) between the mel-cepstral models of the HSMM as a distance metric. Since the mel-cepstral parameters are modeled by k-dimensional multivariate Gaussian distributions and not Gaussian mixture models, we can use the analytic form of KLD (Hershey et al., 2007). By using a symmetric version of KLD, we ensure that the whole interpolation is also symmetric (Müller, 2007).

For each mapping along the warping path, an HSMM state is generated by interpolating the acoustic features and durations of the mapped HSMM states. This sequence of states is then used to synthesize the interpolated utterance. Fig. 1 shows an actual DTW alignment for the beginning of a sentence. We see the optimal warping path that

[3] In GOI, the definite article "die" is reduced to [d̥] and subsequently merged with the initial [d̥] of "Tage".

Table 3
Phone sets used in the experiments, represented with IPA symbols.

Category	(R)SAG	VD	IVG	GOI
Vowels (monoph.)	ɑ ɑː ɒ ɒː ɐ e ɛ eː i ɪ iː ɔ o oː øː æː ɐ œ œː ə u ʊ uː ʏ y yː	ɑ ɑː ɒ ɒː ɐ e ɛ eː i ɪ iː ɔ o oː œː øː æː ɐ œ ə u̯ ʊ uː y yː	ɑ ɑː ɒ ɒː e ɛ eː ɛː i ɪ iː ɔ o ɔː oː ɐ œ ə u ʊ uː ʏ y yː	ɑ ɑː ɒ ɒː e ɛ eː ɛː i ɪ iː ɔ o ɔː oː ø øː ɐ œ u ʊ uː ʏ y yː
Vowels (monoph.) nasalized	ɒ̃ː ɔː æ̃ː œ̃ː	ɒ̃ ɒ̃ː ɑ̃ː ɛ̃ː ĩː õː æ̃ː		ɒ̃ ɒ̃ː ɑ̃ ɑ̃ː ɛ̃ ɛ̃ː ĩː ĩ ɔ̃ ɐ̃ ũ
Diphthongs	aɪ̯ ɒːɐ̯ ɑːɐ̯ ɒɪ̯ ɑʊ̯ ɛɐ̯ ɛːɐ̯ ɪɐ̯ iɐ̯ iːɐ̯ ɔɐ̯ ɔːɐ̯ oːɐ̯ ɔʏ̯ øːɐ̯ œɐ̯ ʊɐ̯ uːɐ̯ oːɐ̯ ʏɐ̯ yːɐ̯	ɒːɐ̯ ɒɪ̯ ɛɐ̯ ɛːɐ̯ iɐ̯ iːɐ̯ ɔɐ̯ ɔːɐ̯ øːɐ̯ œɐ̯ ʊɐ̯ uːɐ̯ oːɐ̯ ʏɐ̯ yːɐ̯	aɐ̯ aɛ̯ ɑɔ̯ ɑɔɐ̯ ɑoɐ̯ eɑ̯ ɛɑ̯ eɛ̯ eɛ̯ ɛɛ̯ ɛi̯ ɛɐ̯ iɛ̯ ɪi̯ iɐ̯ ɔɑ̯ oɑ̯ ɔɛ̯ oɛ̯ ɔɔ̯ ɔɐ̯ ɔɔ̯ ɐɑ̯ ɐo̯ uɪ̯ oɪ̯ ʊu̯	ɑɒ̯ aɐ̯ ɑo̯ ɑɔ̯ eɑ̯ ɛɑ̯ ɛɐ̯ eɐ̯ iɛ̯ iɐ̯ ɔɛ̯ oɪ̯ ɔo̯ ɔɐ̯ ɔu̯ œɛ̯ œɐ̯ uɑ̯ uɐ̯ oɐ̯ uɐ̯ ʊu̯ yɐ̯ ɑɒ̃ aɐ̃ ɑo̯ ɛɐ̃ ɔɑ̃ oɑ̃
Diphthongs nasalized				
Plosives (stops)	ɓ ɖ ɠ k ʔ p t	ɓ ɖ ɠ k ʔ p t	ɓ b ɖ d ɖː ɠ g k kʰ ʔ t tʰ	ɓ ɖ ɠ k kʰ ʔ t tʰ pʰ
Nasal stops	m n ŋ	m n ŋ	m n ŋ	m n ŋ
Fricatives	ç x f h s ʃ v z ʒ	ç x f h s ʃ v	β ä ç x f ɣ h s ʃ v	β ç x f ɣ h s ʃ v
Affricates			t͡ɓf kˣ t͡ks tʃ ts	t͡ɓf t͡ks ts
Approximants	j	j	j	j
Trill	r	r	ʀ ʀˣ	ʀ
Lateral approx.	l	l	l ɭ	l

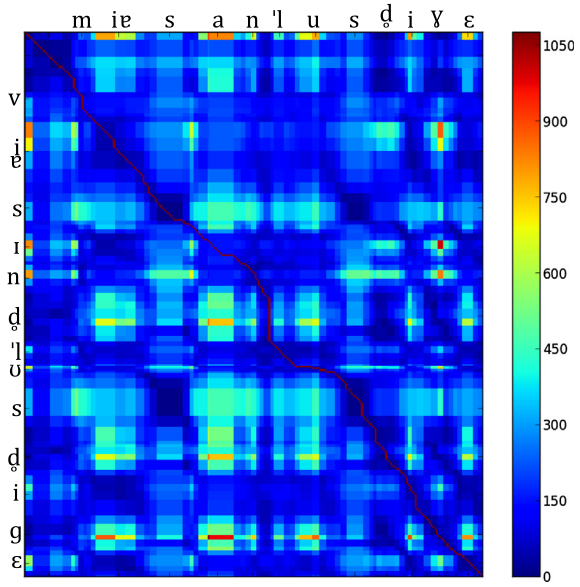


Fig. 1. RSAG-VD interpolation between expanded state sequences of the sequence "Wir sind lustige" (engl. "We are funny").

closely follows the diagonal. Mel-cepstral, $F0$ and aperiodicity features are linearly interpolated between the HSMM state models mapped by the warping path and according to an interpolation parameter α. The interpolation for acoustic features is described in Yoshimura et al. (2000). Horizontal and vertical parts of the warping path indicate one-to-many mappings between states. These type of mappings have to be treated carefully to assure that the total duration of the segment is correct. Our algorithm to handle these cases is described in detail in Section 4.

It should be noted that because the final HSMM is constructed from the DTW warping path, we achieve an approximate interpolation where the endpoints ($\alpha = 0.0$ and $\alpha = 1.0$) are not exactly the original models of RSAG and dialect (GOI, IVG, VD).

Using DTW it is possible to obtain different warping paths in terms of state alignment for unexpanded states where each state has a duration and expanded states where the states are expanded according to the duration of the state. For example, an unexpanded state with duration n would be expanded to n states with duration 1 (the number of feature frames generated from this model). An illustration for this can be seen in Fig. 2 where a 5 state HSMM is shown in unexpanded and expanded form.

An example for the effect of this on the DTW result can be seen in Tables 4 and 5. In this example we use $cost(b,d) = 1 < cost(b,c) = 2 < cost(a,c) = 3 < cost(a,d) = 4$ as a cost function for DTW. With the unexpanded method we can achieve only the mappings $\{a \leftrightarrow c, b \leftrightarrow d\}$, while the expanded method gives us $\{a \leftrightarrow c, b \leftrightarrow c, b \leftrightarrow d\}$ which shows the greater flexibility of the expanded method.

In order to evaluate whether the behavior shown in Table 4 and 5 is actually present in the interpolations of empirical speech samples, we analyzed the alignments of all interpolations of the data set used for our evaluation which we present in Section 7. Fig. 3 shows that there is

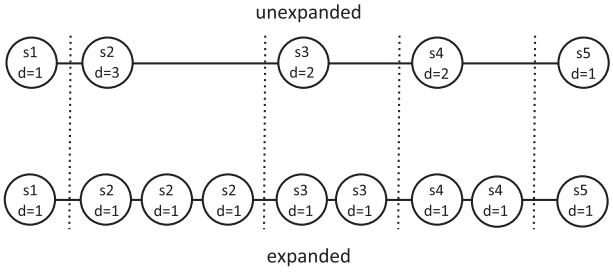


Fig. 2. Illustration of the same HSMM in unexpanded and expanded form. d represents the duration (in feature frames) for each state.

Table 4
Dynamic-time-warping between state sequences "ab" and "cd". State "a" is aligned with "c" and state "b" is aligned with "d" (optimal path in bold).

b	∞	5	**4**
a	∞	**3**	7
	0	∞	∞
		c	d

Table 5
Dynamic-time-warping between expanded state sequence "aabb" and "cccdd". States "a" are aligned with "c", states "d" are aligned with "b" but states "c" are aligned with "a" and "b" (optimal path in bold).

b	∞	10	10	10	9	**10**
b	∞	8	8	**8**	**9**	10
a	∞	6	**6**	9	13	17
a	∞	**3**	6	9	13	17
	0	∞	∞	∞	∞	∞
		c	c	c	d	d

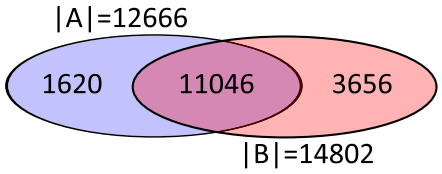


Fig. 3. Counts of unexpanded mappings (A) and expanded mappings (B).

a significant amount of additional unique mappings with the expanded alignment method. There are 3656 mappings between expanded states which do not exist when the unexpanded method is used (e.g. the illustrative mapping b ↔ c in Table 5).

4. Duration modeling

This chapter describes how the durations for the final HSMM, which is constructed from the DTW warping path, are calculated.

For a given utterance, a HSMM state sequence is constructed for each variety involved in the interpolation. HSMMs are retrieved from the voice models by classifying the utterance labels using the voice model decision trees as described by Zen et al. (2009). DTW is then used to calculate an optimal state mapping of the two HSMM state sequences. The result of DTW might contain one-to-many mappings. This is always true when the number of states is different for the two sequences. We have to handle these cases so that the total duration of the final HSMM state sequence is also the result of a linear interpolation between the total durations of the individual sequences with respect to α.

For the standard interpolation (Yoshimura et al., 2000) between random variables, we can use the fact that a linear combination of normally distributed random variables is again normally distributed with mean $a\mu_1 + b\mu_2$ and variance $a^2\sigma^2 + b^2\sigma^2$.

However, for one-to-many mappings we have to interpolate one random variable with multiple other random variables, resulting in a non-linear combination. Consider a mapping of two random variables X, Y with one random variable Z. To interpolate between X, Y and Z, we define the resulting random variables U and V as shown in Eqs. (1) and (2).

$$U = (\alpha X + \alpha Y + (1-\alpha)Z)\frac{X}{X+Y} \quad (1)$$

$$V = (\alpha X + \alpha Y + (1-\alpha)Z)\frac{Y}{X+Y} \quad (2)$$

U and V then model the durations of the two resulting states of the interpolated HSMM sequence. Fig. 4 shows a graphical interpretation of the two Eqs. (1) and (2). In the case of an interpolation value of $1.0, U = X$ and $V = Y$. In the case of an interpolation value of 0.0, the value of Z is distributed on U and V according to the relative values of X and Y. In other words, the duration $X + Y$ is interpolated with duration Z according to α. The resulting duration is then distributed to U and V according to the ratio of X and Y.

To obtain the distribution for Eqs. (1) and (2) in general, we would have to take into account that the product of normally distributed random variables is not normally distributed (Springer and Thompson, 1970) and that the reciprocal of a normally distributed random variable is also not normally distributed.

However, in the synthesis system used here (Zen et al., 2009), the mean values of the duration random variables are used as the actual state durations, as long as no modification of the speaking rate using variance scaling is desired (Valentini-Botinhao et al., 2014). So for the implementation of the interpolation algorithm, we can apply Eqs. (1) and (2) directly on the final duration values d_i (i.e. the means) of the two utterances that we want to interpolate. Duration d_i then specifies the number of feature frames that are generated for state i in the DTW path.

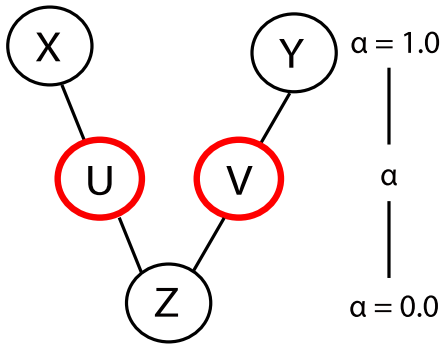


Fig. 4. Interpolation between random variables X, Y, and Z.

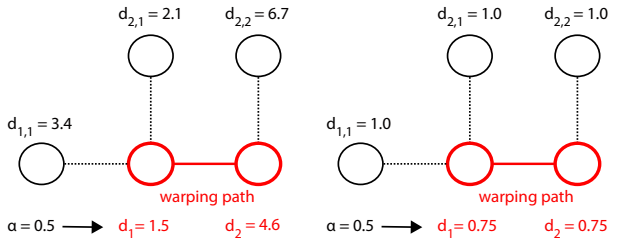


Fig. 5. Interpolation of unexpanded state durations (left) and expanded state durations (right).

This means that for every mapping along the DTW path, a single HSMM state with duration d_i is generated.

For both methods (expanded and unexpanded), we compute the interpolated duration value of the i-th state in the alignment d as shown in Eq. (3). Here $\langle d_{1,1},\ldots,d_{1,m}\rangle$ and $\langle d_{2,1},\ldots,d_{2,n}\rangle$ are the mean duration sequences involved in the interpolation. For a one-to-one mapping this results in $m=n=1$, reducing the formula to standard interpolation. For one-to-many interpolation, we get either $m=1, n>1$ or $m>1, n=1$.

$$d_i=\left(\sum_{j=1}^{m} d_{1,j}\alpha+\sum_{k=1}^{n} d_{2,k}(1-\alpha)\right)\frac{d_{1,i}}{\sum_j d_{1,j}}\frac{d_{2,i}}{\sum_k d_{2,k}}. \tag{3}$$

Fig. 5 illustrates the duration interpolation for a representative case of unexpanded and expanded states with interpolation weight $\alpha=0.5$. For the unexpanded example (left), the mapped sequences that we derive from the DTW alignment are shown in Eqs. (4) and (5). According to the first part of Eq. (3), the total duration of d_1 and d_2 is given by $d_{1,1}\alpha+(d_{2,1}+d_{2,2})(1-\alpha)=(d_1+d_2)=6.1$. The second part of Eq. (3) then distributes the total duration to d_1 and d_2 according to the relation between $d_{2,1}$ and $d_{2,2}$, resulting in $d_1=1.5$ and $d_2=4.6$.

$$\langle d_{2,1}\rangle=\langle 3.4\rangle \tag{4}$$

$$\langle d_{2,1},d_{2,2}\rangle=\langle 2.1,6.7\rangle \tag{5}$$

The example on the right side of Fig. 5 shows an unexpanded case. Here, the interpolated total duration (1.5) is distributed uniformly to the two warping path states $d_1=0.75$ and $d_2=0.75$ as all mapped state durations are 1.0.

Both methods might produce states with durations smaller than 1. To cope with this, we accumulate the durations and skip states according to Algorithm 1. The algorithm loops the final HSMM state sequence. The duration of each state (as calculated before) is accumulated in *accdur*. If *accdur* <1 then the current state is skipped. Else the state is added to the final model and its duration subtracted from *accdur*.

Table 6
Analyzed sentence RSAG–IVG.

SAG orth.	**Schnee liegt im Garten.**
RSAG phones	ˈʃneː liːgd̥ ʔɪm ˈgɑːd̥n̥
IVG phones	ˈʃneə̯ liːt iŋ ˈgɔχtɛ

Algorithm 1. Algorithm for skipping states.

```
accdur ← 0
for all duration, state in HSMM state sequence
  do
    accdur ← accdur + duration
    if accdur ⩾ 1 then
      accdur ← accdur − −duration
      add(state)
    end if
end for
```

The final HSMM state sequence is then used as input for the parameter and waveform generation (Zen et al., 2009). Unsupervised interpolation allows us to generate intermediate variants of utterances for any given utterance pair. In addition, it is also possible to deal with missing words. In terms of linguistic correctness we might produce utterance variants that are wrong in the sense that such intermediate variants do not exist or that co-occurrence[4] requirements are not met.

5. Phonetic analysis of interpolation errors

We applied the previously presented interpolation methods to the sample utterances and dialects as described in Section 2. Here we present an analysis of the input-switch rules and of the processes involved in the interpolation from the (R)SAG input to the dialect output. While all utterances used in the evaluation have been analyzed to extract the necessary information for the extended method presented in Section 6, here we only present one sample utterance per variety. Subsequently, narrow phonetic transcriptions of the diverse interpolated steps from (R)SAG to the respective dialect (IVG; GOI; VD) are shown. 0.0 denotes the (R)SAG synthesis as derived/synthesized from the Standard corpus produced by the respective IVG, GOI, or VD speaker, 1.0 denotes the dialect synthesis. 0.2, 0.4, 0.6, 0.8 are the intermediate forms created by interpolation.

5.1. Interpolation RSAG–IVG

The example sentence analyzed in this section is shown in Table 6.

5.1.1. Input-switch rules

- /ʃneː/ ↔ /ʃneə̯/: A prominent South Bavarian characteristic is the diphthongization of Middle High German (MHG) <ê> in e.g., **Schnee**, (engl. "snow").

[4] Co-occurence requirements refer to the fact that within an utterance, it is not allowed to arbitrarily mix standard and dialect forms (Scheutz, 1999).

Consequently, in IVG, we find [ʃneą] for RSAG [ʃne:]. Since this is a historical process, the alternation from /ʃne:/ ↔ /ʃneą/ cannot be captured as a diphthongization process, but has to be described as an input-switch rule; speakers can realize either [ʃne:] or [ʃneą], but no in-between forms are allowed.

- <-gt> ↔ <-t>: In **liegt**, a similar case is at hand. In IVG and in this area of Eastern Tyrol, the final consonant clusters <-gt> are dissolved (Kranzmayer, 1956). These consonant clusters evolved in the course of contraction processes from Old High German (OHG) **ligit** to New High German (NHG) **liegt** and is already described for MHG. From a synchronic perspective, however, we find either [li:t] which is indicative of a dialectal pronunciation, or [li:gt] which indexes RSAG.
- /ɪ, ʊ/[5] ↔ /i, u/[6]: Phoneme inventories of the Bavarian dialects contain no high lax vowels, therefore, RSAG /ɪm/ alternates with IVG /in/.[7] In the Bavarian dialects, [ɪ, ʊ] might occur as a result of the reduction processes, therefore, on the phonetic level, both tense and lax vowels turn up. Moreover, even in SAG, the distance between high tense and high lax vowels is rather small (Brandstätter and Moosmüller, 2015; Brandstätter et al., 2015), consequently, this input-switch rule is easily feasible in interpolation.
- /ɑ/ ↔ /ɔ/: In many cases, [ɔ(:)] is used instead of SAG [ɑ(:)]. Consequently, the vowel /ɑ/ in SAG **Garten** (engl. "garden") is realized as [ɔ] in IVG.
- <-e>: A further input-switch rule affects the suffix <-en> in **Garten**. In SAG, this suffix is either fully pronounced, resulting in [ɛn], or, in most cases, the vowel is deleted and the remaining nasal consonant is syllabified, resulting in [n̩]. In IVG, however, MHG morphological traces are still apparent: OHG **garto** changed to MHG **garte**. This form is still preserved in IVG.

5.1.2. Phonological processes

Nasal place assimilation is a phonological process present both in RSAG and in the dialects. In IVG, nasal place assimilation of /n/ → [ŋ] takes place in front of the velar plosive [g̊] of Garten. However, nasal place assimilation is not applied in RSAG, because the bilabial nasal consonant /m/ is not subjected to place assimilation in the Austrian varieties. Therefore, a difference between IVG and RSAG is present at a higher, syntactic level which affects the phonological level.

In RSAG **Garten**, /ʀ/ is vocalized. Vocalization is a phonological process which is applied in all Middle Bavarian dialects. Preceding consonants or word-final, /ʀ/ is vocalized to [ɐ̯], following the vowel [ɑ], the result of the vocalization, [ɐ̯], is absorbed and compensatory lengthening of [ɑ] takes place. Therefore, the sequence /ɑʀ/ is pronounced [ɑ:] in SAG. Vocalization of /ʀ/ does not hold for the South Bavarian dialects. Since /ʀ/ is generally not pronounced as a trill, but rather as a fricative, mostly unvoiced, the sequence /ɔʀ/ is pronounced as [ɔχ] (see also Hornung and Roitinger (2000)).

5.1.3. Results of the interpolation steps: phonetic analysis

The analysis of the interpolation output is shown in Table 7 and described subsequently.

Although phonologically, /e:/ to /eą/ in **Schnee** is modeled as an input-switch rule, because only either /e:/ or /eą/ are possible output forms, this input-switch rule can easily be captured as a gradual process of diphthongization in interpolation, as it can be observed in Fig. 6. The first change affects the offset of the vowel. The last 30% of the vowel are marked by a lowering of F2 and F3 in step 0.2. In 0.4, the lowering of F2 and F3 starts even earlier, so that a slight diphthongization [eɛ̯] is perceivable. From now on, F3 stays stable. In 0.6, F2 is lowered in the first part of the vowel, yielding a change in vowel quality from [e] to [ɛ] in the first part. This change is further enhanced in step 0.8 and the final step 1.0 by a substantial raise in F1. At the same time, F2 of the third half of the vowel experiences a further lowering plus a gradual and substantial raise in F1, yielding the vowel qualities [æ] and, finally, [ą]. Tentatively, it can be concluded that after step 0.4, changes are such that they proceed into the direction of the dialect pronunciation.

This assumption is strongly supported by the analysis of the input-switch rule /ɑ/ ↔ /ɔ/ in **Garten**. Fig. 6 shows that F2 trajectories of step 0.0, 0.2, and 0.4 (mean = 1262 Hz) are clearly set apart from the trajectories of steps 0.6, 0.8, and 1.0 (mean = 963 Hz). The difference in F1 is not as straightforward, and the changes from step 0.0 to 1.0 appear more gradual, nevertheless F1 of step 0.0, 0.2, and 0.4 occupies the higher frequency range (mean = 680 Hz), whereas F1 of steps 0.6, 0.8, and 1.0 is located in the lower frequency range (mean = 534 Hz). Consequently, the vowel clearly has the quality [ɑ] in steps

Table 7
Phonetic analysis of unsupervised state-based interpolation between Regional Standard Austrain German (RSAG) and Innervilgraten dialect (IVG).

SAG orth.	**Schnee liegt im Garten.**
RSAG 0.0	ˈʃne: li:gd̥ ʔɪm ˈgɑ:d̥n̩
0.2	ˈʃne: li:gd̥ ʔɪm ˈgɑ:d̥n̩
0.4	ˈʃneɛ̯ li:d̥ ʔɪm ˈgɑˑʁd̥n̩
0.6	ˈʃnɛæ̯ li:d̥$^{\theta}$ɪm ˈgɒχtɛn
0.8	ˈʃnɛæ̯ li:d̥$^{\theta}$ɪŋ ˈgɔχtɛ^{n}
IVG 1.0	ˈʃneą li:t iŋ ˈgɔχtɛ

[5] The vowel inventory of RSAG comprises 13 vowels: /i, ɪ, y, ʏ, e, ɛ, ø, œ, u, ʊ, o, ɔ, ɑ/

[6] The vowel inventory of the Bavarian dialects comprises 7 vowels /i, e, ɛ, u, o, ɔ, ɑ/. As a result of the Viennese monophthongization, two further long vowels have been added for the VD /æ:, ɒ:/.

[7] In the Bavarian dialects, dative and accusative often collapse, therefore in IVG it reads **in Garten** or **in Haus**, while in SAG it is **im Garten** or **im Haus** (engl. "in the garden", "in the house").

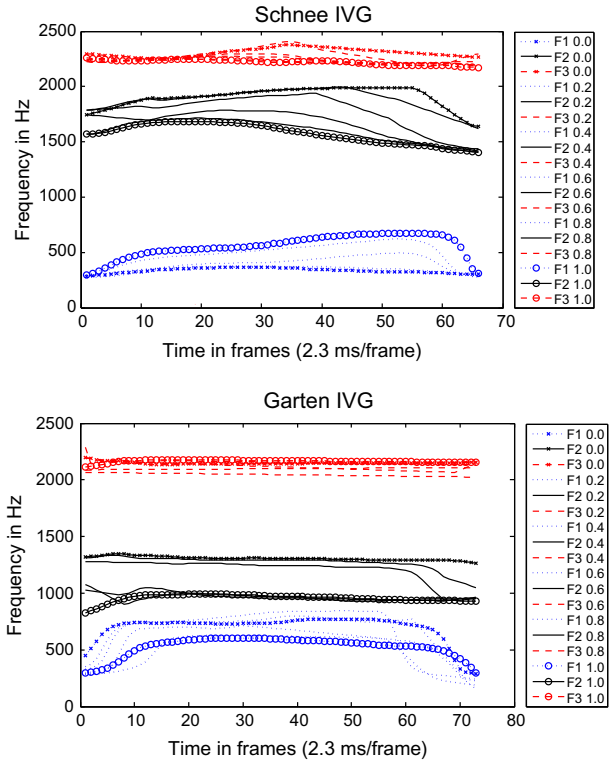


Fig. 6. Formants F1–F3 for RSAG to IVG interpolation of /e:/ from **Schnee** (top) and /ɑ/ from **Garten** (bottom). F1 is shown in dotted lines, F2 solid and F3 dashed. $\alpha = 0.0$ uses crosses and $\alpha = 1.0$ uses circles as line markers to represent the interpolation endpoints. Formant trajectories are linearly time aligned in order to facilitate comparison of changes over time, therefore, changes in duration are not obvious from the figures.

0.0, 0.2, and 0.4, whereas a quality change from [ɑ] to [ɒ] starts in step 0.6, which is finally changed to [ɔ] in steps 0.8 and 1.0. Voicing of the velar stop /g̊/ of **Garten** starts in 0.2. Voicing of stops preceding nasal consonants is a general process in all varieties of Austrian German, though restricted to word-internal positions in SAG.

In SAG and in the Middle Bavarian varieties in general, /ʁ/ is vocalized. This is not the case in the South Bavarian varieties. Therefore, the IVG sequence /ɔʁ/ is pronounced [ɔχ]. In interpolation from [ɑː] to [ɔχ], [χ] has to be inserted. Insertion starts in step 0.4. First, the final part of the long vowel [ɑː] is fricatized, the fricative [ʁ] is still short and voiced, and therefore, the vowel is only slightly shortened to [ɑˑ]. In step 0.6, the fricative is devoiced, simultaneously, fortition of [d̥] takes place. The suffix <-en>, which reads <-e> in IVG, still needs to be changed. In 0.6, [ɛ] is inserted, leading to the sequence [ˈgɒᵡtɛn], however, this output sequence of **Garten** violates co-occurrence restrictions, since the nasal consonant is still fully pronounced. Deletion of the nasal only starts in 0.8, together with the full pronunciation of the fricative [χ]. Nasal assimilation of /n/ → [ŋ] is accomplished in 0.8.

The critical steps in the derivation are to be found in 0.4 and in 0.6. 0.6 is affected the most. Whilst 0.4 might still be evaluated as SAG, 0.6 can neither be assigned to SAG nor to IVG, either because the deviation of the intermediate steps is too large from both SAG and IVG, as, e.g., in **Schnee** [ˈʃnɛæ̯], or because co-occurrence restrictions are violated, as in **Garten** [ˈgɒᵡtɛn]. Nasal assimilation of IVG /n/ → [ŋ] and the deletion of the final nasal consonant in **Garten** should start one step earlier.

5.2. Interpolation RSAG–GOI

The example sentence analyzed in this section is shown in Table 8.

5.2.1. Input-switch rules

- /kɔmt/ ↔ /kimd̥/ **kommt**: In the dialects, the root vowel of the verb **kommen** (engl. "to come") is either /e, i/ or /u/. In GOI, the first variant is used, leading to the following inflections: /kim, kimd̥, kimsd̥, kemɑn, kemd̥s, kemɑnd̥/.
- [maɐ̯nɛ] ↔ [mãẽ̯] **meine**: Final nasals have been deleted in Bavarian dialects and the preceding vowel/diphthong has been nasalized. Therefore, we find [mãẽ̯] in GOI, opposed to RSAG [maɐ̯nɛ]. Again, historically, these are phonological processes, yet, in the synchronic view, no intermediate steps can be observed.
- /ɛ/ ↔ /e/ in **Schwester**: The pronunciation of the e-vowels is often reversed. Therefore, we find /ɛ/ in SAG, but /e/ in GOI.
- /p, t/ ↔ /b̥, d̥/: Front plosives are neutralized in the Bavarian dialects, as, e.g., in GOI [ˈʃvesd̥ɐ].

5.2.2. Phonological processes

In **morgen** (engl. "tomorrow"), different phonological inputs have to be assumed. However, phonological processes are quite similar. In SAG, /ˈmɔʁgɛn/ is assumed. /ʁ/ is vocalized to [ɐ̯], resulting in [ˈmɔɐ̯gŋ̍] in (R)SAG. The deletion of unstressed [ɛ] with subsequent nasal place assimilation finally results in [ˈmɔɐ̯gn] or even [ˈmɔɐ̯ŋ̍].

For GOI, /ˈmɔʁigɛn/ has to be assumed, since the insertion of an epenthetic vowel which splits the clusters /ʁ/ plus velar or labial obstruents goes back to the 13th century and is retained in some persistent dialects (Kranzmayer, 1956). In the same way as in SAG, deletion of the unstressed vowel [ɛ] with subsequent nasal place assimilation takes place. Subsequently and unlike SAG, the velar plosive is

Table 8
Analyzed sentence RSAG–GOI.

SAG orth.	**Morgen kommt meine Schwester.**
RSAG phon.	ˈmɔɐ̯gŋ̍ kɔmt maɐ̯nɛ ˈʃvᵉestɐ
GOI phon.	ˈmɔʁɪŋ̍ kimd̥ mãẽ̯ ˈʃvᵉesd̥a

Table 9
Phonetic analysis of unsupervised state-based interpolation between Regional Standard Austrian German (RSAG) and Bad Goisern dialect (GOI).

SAG orth.	**Morgen kommt meine Schwester.**
RSAG 0.0	ˈmɔɐ̯gŋ̍ kɔmd̥ maɐ̯nɛ ˈʃvᵉesd̥ɐ
0.2	ˈmɔɐ̯gŋ̍ kʊ̃md̥ maɐ̯ne ˈʃvᵉesd̥ɐ
0.4	ˈmɔˑᴮʀɪŋ̍ kʊ̃md̥ maɐ̯nᵉ ˈʃvᴱesd̥ɐ
0.6	ˈmɔːʀɪŋ̍ kimd̥ maɐ̯̃ː ˈʃvᴱesd̥ɐ
0.8	ˈmɔːʀɪŋ̍ kˣimd̥ maɐ̯̃ː ˈʃvᴱesd̥a
GOI 1.0	ˈmɔːʀɪŋ̍ kˣimd̥ maɐ̯̃ː ˈʃvᴱesd̥a

deleted, leading to the output [ˈɔʀɪŋ̍].[8] Vocalization of /ʀ/ is applied in **Schwester** (engl. "sister") in both SAG and GOI.

5.2.3. Results of the interpolation steps: Phonetic analysis

The analysis of the interpolation output is shown in Table 9 and described subsequently.

As outlined in Section 5.2.2, for **morgen** (engl. "tomorrow"), two different forms have to be assumed phonologically. However, the processes involved can be easily captured by interpolation. As a first step, (R)SAG [ɔɐ̯] has to be monophthongized to [ɔ]. Monophthongization only affects the offset of the diphthong, since r-vocalization needs to be undone. From step 0.0 to 0.4 a gradual lowering of F2 at the offset of [ɔɐ̯] is visible in Fig. 7. In step 0.6, monophthongization is accomplished with respect to F2. A slight movement of F1, equal to step 0.4, is still distinguishable. Also, F3 of steps 0.4 and 0.6 lies inbetween the RSAG and the dialectal form. Step 0.4 is the most crucial one; a slight diphthong is still audible, but simultaneously, the velar stop has been changed to a fricative and the vowel [ɪ] has been inserted, which changes its quality to [i] in step 0.8.

The input-switch rule /ɔ/ ↔ /i/ in **kommt** (engl. "come", 3rd sg.), involves dramatic changes which predominantly affect F2, since a relatively low F2 for the back vowel [ɔ] has to be changed to a high F2 for the front vowel [i] (see Fig. 7). However, F1 and F3 are affected as well: F1 has to be lowered and F3 has to be raised in order to attain [i]. Steps 0.0 and 0.2 are quite similar with respect to formant trajectories, yet, the auditive impression changes a bit, maybe due to a decrease in duration.

Steps 0.4 and 0.6 are severely impaired; in step 0.4, F2 could not be fully detected, and the auditive result is not reliably assignable to any vowel, apart from its being heavily nasalized. The transcription offered has therefore to be understood as a compromise. F1 of step 0.4 is rather low (mean = 282 Hz), thus suggesting [i]. F3, however, is rather low as well (mean = 2536 Hz), thus indicating [ɔ]. In step 0.6, F3 could not fully be detected, however, it is rather high (mean = 2718 Hz) and suggests to proceed in the same way as steps 0.8 and 1.0. Therefore, F3 indicates [i]-quality, along with a high F2 and a low F1. The auditive result is clearly [i]. The quality of the vowel [i] is fully accomplished in steps 0.8 and 1.0.

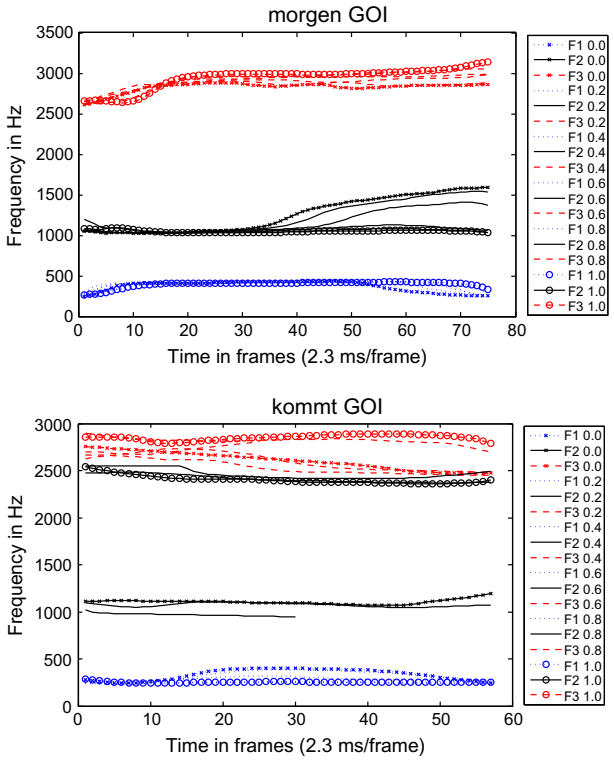


Fig. 7. Formants F1–F3 for RSAG to GOI interpolation of [ɔɐ̯] from **morgen** (top) and /ɔ/ from **kommt** (bottom).

The diphthongization of **Schwester** (engl. "sister") is smoothly captured by interpolation (see Fig. 8). As discussed in Section 2, the speaker realizes a rather regional variety of SAG which demands [ɛ]. Our speaker, however, starts off with a slightly diphthongized vowel [ᵉe]. The first part, which is responsible for the diphthongization, takes up the same time span over all interpolation steps, however, from step to step, diphthongization becomes more pronounced. F1 and F3 play no role, most probably, because the starting point already strongly resembles the endpoint.

In the Bavarian dialects, unstressed syllables are generally not reduced to a schwa-like vowel, but the full vowel quality is preserved. This is obvious in the final syllable <-er> of **Schwester** (engl. "sister"). In interpolation, this process is observable by the changes of F1 which is gradually raised (see Fig. 8). Raising of F1 starts at 0.4 and affects approximately the first third of the vowel. It stays constant in step 0.6, but both step 0.4 and 0.6 already give the auditive impression of a full vowel. In step 0.8 and 1.0, the raise of F1 affects most of the vowel, thus indicating the full vowel quality.

5.3. Interpolation RSAG–VD

The example sentence analyzed in this section is shown in Table 10.

[8] Deletion of the stop takes place in all dialects, in VD, e.g., the phonetic output would read [ˈmɔɐ̯ŋ̍].

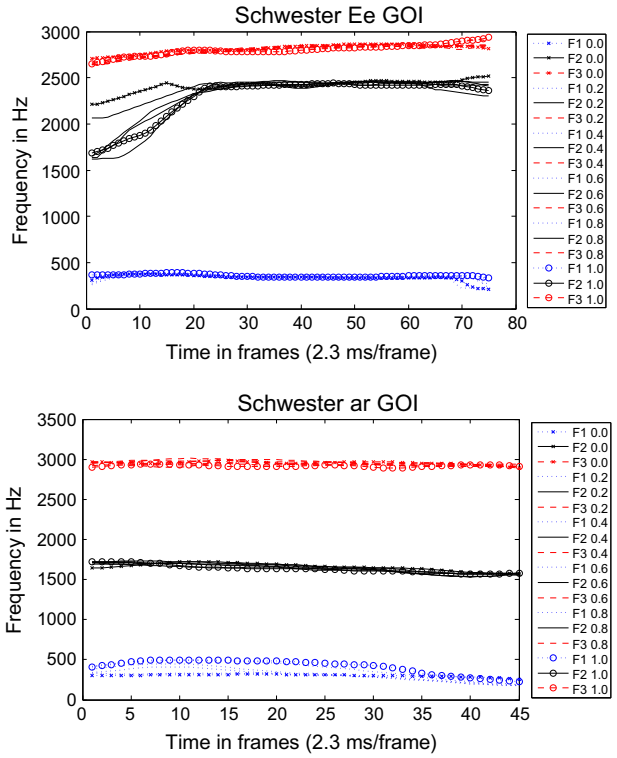


Fig. 8. Formants F1–F3 for RSAG to GOI interpolation of [ᵉe] from **Schwester** (top) and <-er> from **Schwester** (bottom).

Table 10
Analyzed sentence RSAG–VD.

SAG orth.	**Wir sind lustige Leute.**
RSAG phones	ˈviɐ̯ sɪnd̥ ˈlʊsd̥ɪgɛ ˈlɒɛ̯tɛ
VD phones	ˈmɪɐ̯ san ˈɬᵍusd̥ige ˈɬæːtʰ

5.3.1. Input-switch rules

- /viʁ/ ↔ /miʁ/: Alternation of the initial consonant in **wir**: /viʁ/ ↔ /miʁ/. The bilabial nasal consonant is used in all Bavarian dialects of Austria.
- /sɪnd/ ↔ /sɑn/: The alternations of the verb **sein** (engl. "to be") are manifold, in the VD, we find mostly /sɑn/ for the plural.
- /ʊ/ ↔ /u/: As has already been described for IVG, phoneme inventories of the Bavarian dialects contain no high, lax vowels. Consequently, in **lustige** (engl. "funny"), VD /u/ is opposed to SAG /ʊ/.
- /ɔɛ̯/ ↔ /æː/: The vowel inventory of the Bavarian dialects contains no front, labial vowels. This holds also for the diphthong /ɔɛ̯/ which is delabialized to /aɛ̯/. Therefore, in the Bavarian dialects, we find [laɛ̯d̥] with different degrees of diphthongization.
- /aɛ̯, ɑɔ̯/ ↔ /æː, ɒː/: In VD, the diphthongs /aɛ̯/ and /ɑɔ̯/ have been monophthongized to /æː/ and /ɒː/, respectively. Consequently, in the VD, the phonetic output for **Leute** is [ɬæːd̥].

Table 11
Phonetic analysis of unsupervised state-based interpolation between Regional Standard Austrian German (RSAG) and Viennese dialect (VD).

SAG orth.	**Wir sind lustige Leute.**
RSAG 0.0	ˈviɐ̯ sɪnd̥ ˈlʊsd̥ɪgɛ ˈlɒɛ̯tɛ
0.2	ˈviɐ̯ sɪnd̥ ˈlʊsd̥ɪgɛ ˈlɒɛ̯tᵋ
0.4	ˈviɐ̯ sɪ̃nd̥ ˈʊŋsd̥ɪgɛ ˈɒɛ̯tᵋ
0.6	ˈviɐ̯ sɐ̃nd̥ ˈʊŋᵘsd̥ɪgɛ ˈɬæːᵋtʰ
0.8	ˈβiɐ̯ sãn ˈɬᵍusd̥ige ˈɬæːtʰ
VD 1.0	ˈmɪɐ̯ san ˈɬᵍusd̥ige ˈɬæːtʰ

5.3.2. Phonological processes

For both the Bavarian dialects and SAG, r-vocalization applies in /viʁ/ and /miʁ/, resulting in [viɐ̯] and [miɑ̯] or [miɐ̯], respectively.

In **lustige** (engl. "funny"), spirantization of the intervocalic velar plosive takes place in VD, resulting in [ˈlusd̥iɣɛ]. This process might even occur in SAG, especially in spontaneous speech and in high frequency words.

5.3.3. Results of the interpolation steps: Phonetic analysis

The analysis of the interpolation output is shown in Table 11 and described subsequently.

The input switch rule /ɪ/ ↔ /a/ in **sind** (engl. "are") involves a similar dramatic change as has been described for the input-switch rule /ɔ/ ↔ /i/ in GOI, with the difference that now a front vowel has to be transformed into a back vowel.[9]

The changes to be performed from /ɪ/ to /a/ involve a substantial lowering of F2, a lowering of F3, and a substantial raise of F1 (see Fig. 9). The outputs of steps 0.0 and 0.2 contain no changes, the first change starts at step 0.4 which reveals a rather weak, but substantially lowered F2. The quality of [ɪ] is still preserved in F1 and F3, consequently, the auditive impression resembles a heavily nasalized [ɪ̃]. In step 0.6, F3 is lowered as well, but F1 has not changed yet. This enhances the a-quality, which is transcribed as [ɐ̃]. It is only in step 0.8 that a substantial part of F1 is raised, thus producing the quality of a full vowel [ã] which is still nasalized due to the low F1 in the final third of the vowel. Finally, in step 1.0, the whole vowel shows a raised F1.

As outlined in Section 5.3.1, the SAG diphthong /ɔɛ̯/ of **Leute** (engl. "people"), realized/synthesized as [ɒɛ̯] in 0.0, does not belong to the phoneme inventory of the Bavarian dialects. In the Bavarian dialects, the first part of the diphthong is delabialized, resulting – depending on the dialect – in [aɹ̯], [aɛ̯], or [aɐ̯]. In the Viennese dialect, the diphthong is additionally monophthongized, resulting in [æː]. Derounding starts in step 0.4, by a slight raise of F2 and F3. A pronounced diphthongal movement is still evident at step 0.4. The de-rounded diphthong already indicates a dialectal pronunciation, however, some co-

[9] /a/ is a back vowel with respect to the location of constriction (pharyngeal) and a front vowel with respect to the highest point of the tongue (Fant, 1965).

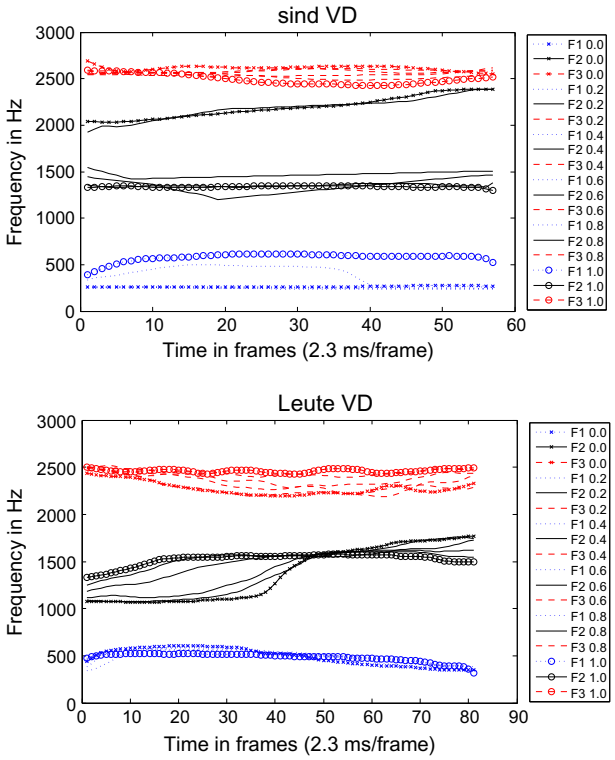


Fig. 9. Formants F1–F3 for RSAG to VD interpolation of /ɪ/ from **sind** (top) and [ɒɛ̯] from **Leute** (bottom).

occurrence restrictions are violated. First of all, the final [ɛ] of Leute is not yet fully deleted. Then, the lateral is already velarized. The velarized variant of the lateral is restricted to the area of Vienna, therefore, it has to co-occur with a monophthongized [æː]. In step 0.6, the final vowel [ɛ] has been deleted, but a diphthongal movement is still present in the diphthong which needs to be monophthongized, although F2 of the first part of the diphthong has been substantially raised. In steps 0.8 and 1.0, monophthongization is accomplished, the slight raise in F2 in the first 15 frames is due to the transition from the velarized lateral into the vowel.

6. Phonological model for interpolation

Based on the results of the phonetic analysis in Section 5, we extended our interpolation algorithm to handle input-switch rules for those parts the of the utterances for which interpolation is not phonetically feasible.

6.1. Region definition

To incorporate input-switch rules, we add meta information to each pair of utterances (A, B) to be interpolated. For utterance A and B, we define a set of regions $R(A)$ and $R(B)$ on a phone level. Every region $a \in R(A)$ and $b \in R(B)$ has to consist of at least one phone and can, at maximum, span the whole utterance. Also, a region must not necessarily consist only of consecutive phones (i.e. the region can be split up across the region) but the ordering of the phones must be preserved. The regions have to be selected so that a bijection $M : R(A) \mapsto R(B)$ can be defined. This means, every utterance is split into regions and these regions are then mapped between the utterances. For every region mapping $m = (a, M(a)) \forall a \in R(A)$, a procedure to be applied during the interpolation process can be defined. For our experiments, we set the procedure for every mapping in the evaluation data to either "feature interpolation" or "feature switch". Feature interpolation is used in case of a phonological process or a pseudo-phonological process as described in Section 1. We use the feature switch procedure in case of an input-switch rule. Both procedures are described in Section 3. If each utterance forms a single region and the mapping between these two regions is associated with the feature interpolation procedure, we get the basic interpolation method as described previously. To summarize, regions define which procedure from Section 1 should be used for this part of the utterance.

For our experiments, we defined the mappings M and the associated procedures according to the results of our phonetic analysis as follows: From the evaluation data, extract a list of word mappings with an input-switch rule that cannot be modeled by interpolation. If such a word mapping occurs in a sentence, compare the prefix and suffix of the words on a phonological level. If the phone symbols are the same, a region for feature interpolation can be formed. This is useful as the acoustic realization of the phonetic symbols will still differ slightly for all dialects. The remaining, differing phones form a region that will use the feature switch procedure. Finally, merge regions next to each other that have the same procedure assigned to them. A more sophisticated algorithm could involve DTW on the phonological level, combined with a list of phone level mappings that should be realized using a feature switch procedure. For mapping on the word level, machine translation methods (Neubarth et al., 2013) could be employed.

An example for a defined region mapping and procedures can be seen in Fig. 10. As described previously, /viɐ̯/ and /miɐ̯/ are connected by an input-switch rule. As the suffix /iɐ/ is the same in both utterances, it can form the feature interpolation region 2. /v/ and /m/, on the other

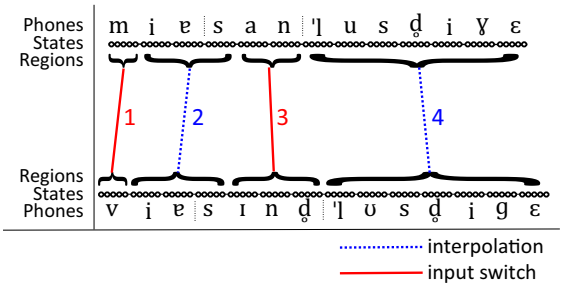


Fig. 10. Example region definition of the sequence "Wir sind lustige" (engl. "We are funny").

hand, form the feature switch region 1. It can also be seen that region 2 is merged with the /s/ from the following word, because this is again a word beginning with the same phones and also uses feature interpolation.

6.2. Region procedures

The function I_{linear} applies DTW on the HSMM states and linearly interpolates the associated features as well as the durations as described in Section 3 and returns the newly generated states.

The function I_{switch} does a feature switch for given HMM states and is shown by Eq. (6).

$$I_{switch}(a,b,\alpha) = \begin{cases} a & \alpha \leqslant 0.5 \\ b & \alpha > 0.5 \end{cases} \tag{6}$$

6.3. Region-based interpolation

The inputs for the extended, region-based interpolation algorithm are: phonetic transcriptions of two utterances, the two associated voice models, the interpolation parameter α and the region information including region mapping and region procedure for each mapping. Algorithm 2 presents the subsequent steps. First, for each region, the indices of the HSMM states representing each phone in this region are retrieved using the voice model decision trees. $value(index)$ is then used to access the actual HSMM state model for the given index. *is_switch* and *is_interpolation* are functions that return true if the supplied region has to use the feature switch or feature interpolation procedure respectively. In case of feature switch, the function I_{switch} is used to retrieve the resulting indices for the current region and is then, together with the associated values, appended to the *results* list. If the region has to be interpolated, *DTW* is applied, which returns a list of tuples of indices, representing the optimal warping path (as described in Section 3). All HSMM states along this warping path are then interpolated using I_{linear} and the resulting, new HSMM states are returned. These are, together with the associated indices for each utterance, also appended to the *results* list. Finally, *results* is sorted according to a function *ordering* (described in 6.4) which defines if the states should be in order of utterance A or utterance B.

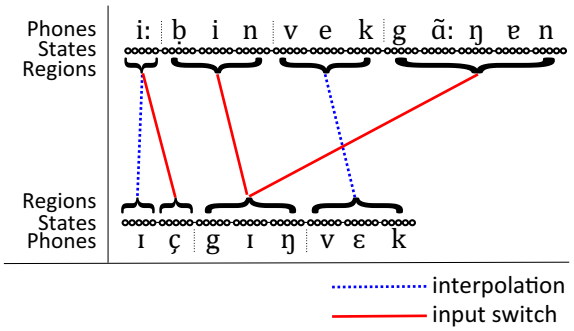


Fig. 11. Example region definition of the translated sentence "Ich ging weg" ↔ "Ich bin weg gegangen".

Algorithm 2. Region-based interpolation algorithm.

1: $results \leftarrow list()$ {result list}
2: **for all** $r \in R(A)$ **do**
3: $idx_A \leftarrow$ indices of HMM states in r
4: $idx_B \leftarrow$ indices of HMM states in $M(r)$
5: **if** is_switch(r) **then**
6: $values = I_{switch}(value(idx_A), value(idx_B), \alpha)$
7: append $(idx_A, idx_B, values)$ to *results*
8: **else if** is_interpolation(r) **then**
9: $dtwpath \leftarrow DTW(idx_A, idx_B)$
10: $values = I_{linear}(dtwpath, \alpha)$ {dtwpath column 0 is indices for first utterance, column 1 for second utterance}
11: append $(dtwpath[0], dtwpath[1], values)$ to *results*
12: **end if**
13: sort *results* by $results[I_{order}()]$
14: **end for**

6.4. Duration and order modeling

In the extended method, duration modeling for feature interpolation is as described in Section 4. For feature switching, this method is not necessary, the number of states and their durations can just be taken from one of the two involved utterances depending on α.

This extended interpolation method can also be used for utterances with a different syntactic structure. Consider the example of a translation from Standard German into SAG syntax that can be seen in Fig. 11: "Ich ging weg" (engl. "I left") ↔ "Ich bin weg gegangen" (engl. "I have left").

In this case, the region definition is a bit different because there are no neighbored regions that could be merged. The region-based interpolation algorithm would then again apply the associated procedures (feature interpolation or feature switch) on each mapped region. Feature interpolation creates states for each mapping along the DTW path. Feature switch uses states, durations and features from either utterance A or from B, depending on α. The states of the regions are then concatenated in the order of A or B, also depending on α. So if $\alpha \leqslant 0.5$, the ordering of A is used, else ordering of B. For the example shown in Fig. 11 this means that for $\alpha \leqslant 0.5$ the ordering of "Ich ging weg" is used, for $\alpha > 0.5$ the ordering of "Ich bin weg gegangen". The evaluations presented in this study did not include utterances which required reordering.

7. Evaluation

We conducted two subjective listening tests to evaluate the presented methods. In the first evaluation we compared the intelligibility and the applicability of interpolation for

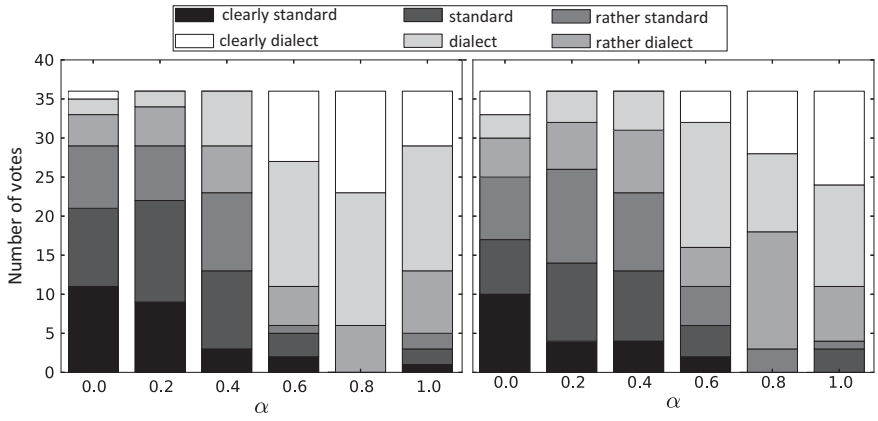


Fig. 12. Scores for unexpanded (left) and expanded (right) method.

the expanded and the unexpanded method. Word-error-rate experiments were carried out in order to evaluate whether the interpolated samples have a higher word-error-rate than the uninterpolated samples, but were not meant to measure the inherent word-error-rates of dialects. The second evaluation was used to assess the effect of the integration of input-switch rules in the interpolation process. Both evaluations are based on the data and training methods described in Section 2.

7.1. Expanded and unexpanded method

In this subjective evaluation we interpolated from RSAG to IVG, RSAG to GOI and RSAG to VD. We used 6 utterances and one speaker per dialect and interpolated each utterance using 6 different values for α (0.0, 0.2, 0.4, 0.6, 0.8, 1.0), using both expanded and unexpanded states. This setup produced 216 unique sound samples for the subjective listening test.

We had 12 native Austrian listeners from age 25 to 64, 4 female and 8 male, who took part in the evaluation. While the listeners were mainly acquainted with VD, 2 listeners were IVG speakers and another 2 listeners grew up near regions where GOI is spoken. The evaluation was split into two parts. The first part was an intelligibility test where the listeners had to write the perceived content of audio samples into a text field. The samples were randomly assigned to the listeners under the constraint that each utterance was heard only once by each listener in order not to bias the evaluation. In the second part, the listeners had to score randomly assigned audio samples according to rate the dialect. The 6 possible score levels were: "Clearly Standard (1)", "Standard (2)", "Rather Standard (3)", "Rather dialect (4)", "Dialect (5)", "Clearly dialect (6)".

Fig. 12 shows the scores for degree of dialect. It can also be seen that the subjective scores strongly changed from standard to dialect from $\alpha = 0.4$ to $\alpha = 0.6$. This seems like a natural boundary for a switch from "rather standard" to "rather dialect" and is actually reflected in the evaluation data. The scores are not significantly different although the unexpanded method shows a slightly higher variation in its mean score.

Table 12
Word-Error-Rates [%] per α and method.

	0.0	0.2	0.4	0.6	0.8	1.0	μ
Unexp.	2.2	6.5	6.5	18.3	8.6	17.2	9.9
Exp.	5.1	7.5	8.6	12.9	28.0	23.7	14.3

The word-error-rate results of the intelligibility test can be seen in Table 12. The unexpanded method has a lower error in all cases except for $\alpha = 0.6$. The word-error-rate of the intermediate variants was not significantly higher ($p > 0.4$) than the full dialectal case ($\alpha = 1.0$) in the Matched Pairs Sentence-Segment Word Error test (Gillick and Cox, 1989). Actually, it was significantly lower ($p < 0.02$) for $\alpha = 0.4$ than for $\alpha = 1.0$. This means that our interpolation methods do not produce a large number of errors, which would result in a higher word-error-rate for the intermediate variants. Although the dialects differ in their prevalence (i.e. VD is understood by many more Austrian inhabitants than IVG), the error rates were not significantly different (IVG: 11.2, GOI: 15.4, VD: 11.1).

While for $\alpha = 0.8$ we saw a large difference in word-error-rate, the overall difference between expanded and unexpanded was not significant. We chose the unexpanded method for our further experiments since it is computationally less expensive.

7.2. Interpolation with and without input-switch rules

For our second subjective evaluation we interpolated from SAG to IVG, GOI and VD. For each of the 3 dialects, using 6 different values for α (0.0, 0.2, 0.4, 0.6, 0.8, 1.0), we interpolated 6 different utterances. Additionally, we generated the same set of samples again, this time including input-switch rules. This setup also produced 216 unique sound samples for the subjective listening test.[10]

[10] Samples at http://userver.ftw.at/~mtoman/specom14/.

34 native dialect listeners (born in the region where the dialect is spoken and raised in the respective dialect, 12 for VD, 11 for IVG, 11 for GOI) from age 13 to 69, 15 female and 19 male, conducted the evaluation.[11] The participants were carefully selected according to their dialect proficiency. Each listener had to score samples for her/his native dialect only. In this evaluation only native dialect listeners participated, since listeners had to answer questions that require a strong knowledge of the respective dialect.

The evaluation consisted of three parts: an intelligibility test, degree of dialect assignment, and a standard/dialect acceptance rating. For the intelligibility test, listeners had to write the perceived content of audio samples into a text field. The samples were randomly assigned to the listeners under the constraint that each utterance was heard only once by each listener. In the second part, each listener had to score all 72 audio samples of his/her dialect (in random order) according to degree of dialect. Since in this second evaluation only native dialect listeners participated, we allowed a more fine-grained control for the degree of dialect using a slider offered by the evaluation user interface. The two ends of the slider were named "standard" and "dialect" respectively, using the name of the actual dialect of the sample. In the third part, listeners had to answer if they accept the speaker of the sample as

- "speaking standard language and grown up with it",
- "speaking standard language which was acquired later in life",
- "speaking dialect and grown up with it",
- "speaking dialect which was acquired later in life",
- or "speaks neither standard nor dialect".

Again we used the name of the actual dialect in these questions, so e.g. "speaking Viennese dialect, grown up with it". In this way we wanted to evaluate the acceptability of our generated varieties.

Fig. 13 shows the median scores for degree of dialect. The largest increase in subjective score again occurred from $\alpha = 0.4$ to $\alpha = 0.6$. As expected, the extended method that handles input-switch rules exhibits a steeper degree change here. For both methods, the subjective rating of the listeners roughly reflects the actual used value for α. This suggests that (linear) interpolation is a reasonable approach for generating in-between variants.

Fig. 14 shows the results of the standard/dialect acceptance test for interpolation without and interpolation with input-switch rules as stacked bar plots respectively. Again it can be seen that a higher α also results in the listeners perceiving the speaker of the samples as being an increasingly more authentic dialect speaker.

Fig. 15 shows the counts of the user choices of "speaks neither standard nor dialect". Here it can be seen that using the method incorporating input-switch rules yields less votes for this category, especially for the strongly interpolated variants at $\alpha = 0.4$ and $\alpha = 0.6$.

Table 13 shows the Word-Error-Rates from the intelligibility part of the evaluation for interpolation only, interpolation with input-switch rules and for the different dialects.

[11] In the future, we plan to repeat the evaluation with dialectologists and linguists.

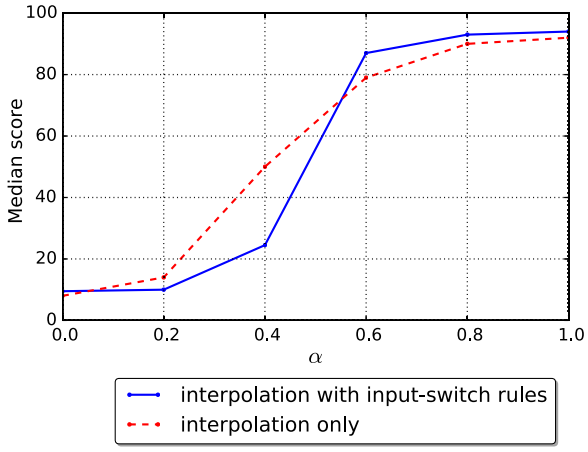


Fig. 13. Median scores for degree of dialect.

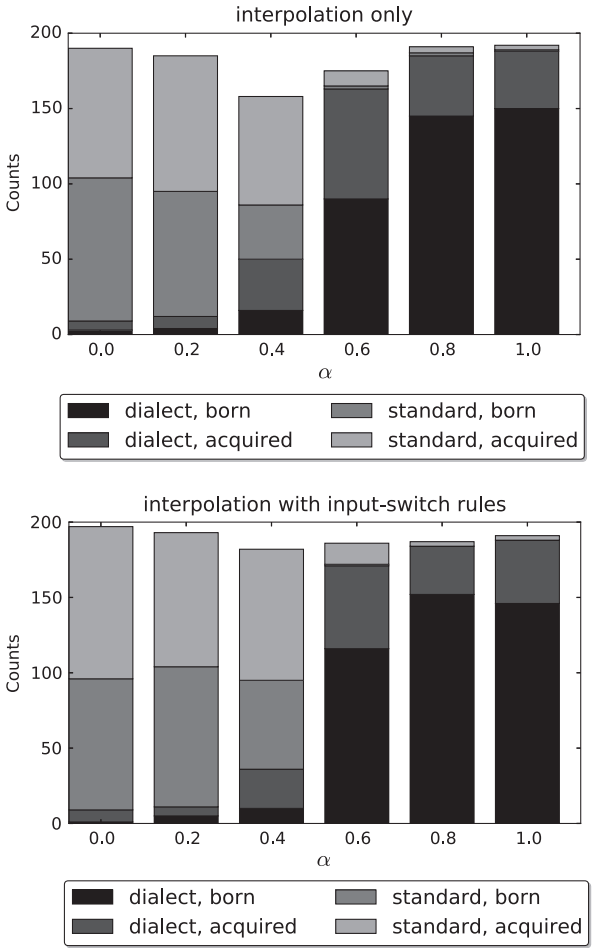


Fig. 14. Speaker category choices.

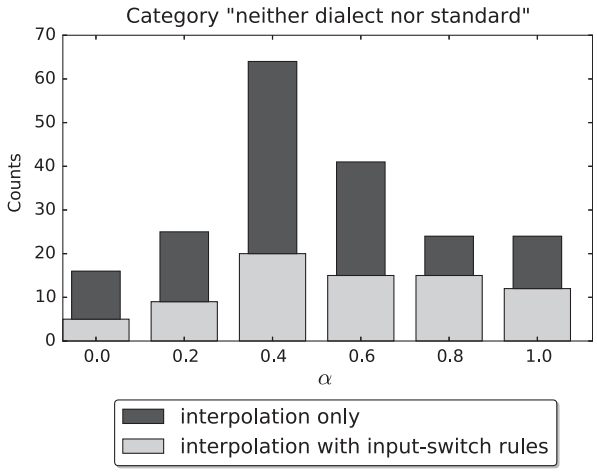


Fig. 15. Category counts for "speaks neither standard nor dialect".

Table 13
Word-Error-Rates [%] for "interpolation only" and "interpolation with switching rules" for the three dialects.

	ipol.	ipol. + switch	IVG	GOI	VD
err.	14.6	13.1	13.9	14.2	13.2

It can be seen that there are no significant differences for the different dialects and also between interpolation and interpolation with input-switch rules. This shows that interpolation by itself does not produce unintelligible intermediate variants and adding input-switch rules does not significantly increase intelligibility but increases the acceptance of the samples as an authentic dialect.

8. Discussion and conclusion

We have presented an unsupervised interpolation method to generate in-between variants of language varieties. It employs DTW to find mappings of HSMM-states for state-level interpolation. Two methods were introduced to handle state durations – either expand each state with duration N to N states with duration 1 (i.e. each state generates 1 feature frame), or continue with the unexpanded states. The results of our experiments suggest that linear interpolation of HSMM-states mapped by DTW is reasonable. Employing DTW on unexpanded or expanded state sequences showed no significant difference, so the computationally less expensive unexpanded method should be preferred. We also presented a method to interpolate state durations for one-to-many mappings.

Based on synthesized samples using these methods, we performed a phonetic analysis to identify utterance sections for which interpolation is not phonetically feasible. The extended method presented here introduces region definitions and region mappings for utterances. The features of the HSMM states of these regions are then subjected to either feature interpolation or feature switch as defined by the results from this analysis.

As expected, this introduces a sudden change in dialect degree perception from values below to above $\alpha = 0.5$, but at the same time decreases the generation of speech that the listeners rated as "neither standard nor dialect". Consequently, for producing correct in-between variants, including input-switch rules is beneficial. Still, even without treating input-switch rules, word error rate did not significantly decrease for highly interpolated samples (α between 0.4 and 0.6) as compared to the pure dialectal samples ($\alpha = 1.0$). Thus, even if no phonetic knowledge about these rules exist, the presented interpolation method still produces intelligible speech.

Future work will include adaptation of the interpolation function so that the subjective perception has a stronger linear correlation with the interpolation parameter α. For example, a piecewise linear function could be employed. A full interpolation system would also employ a machine translation component that translates a standard variety into dialect. The output of this component, which can also include syntactic changes, would provide us with input-switch rules for words. Rules for phonemes have to be derived from phonetic criteria.

Acknowledgments

This work was supported by the Austrian Science Fund (FWF) – Austria: P22890-N23. The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH is funded within the program COMET – Competence Centers for Excellent Technologies by BMVIT, BMWA, and the City of Vienna. The COMET program is managed by the FFG.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.specom.2015.06.005.

References

Astrinaki, M., Yamagishi, J., King, S., D'Alessandro, N., Dutoit, T., 2013. Reactive accent interpolation through an interactive map application. In: Bimbot, F., Cerisara, C., Fougeron, C., Gravier, tG., Lamel, L., Pellegrino, F., Perrier, P. (Eds.), Proceedings of the 14th Conference of the International Speech Communication Association (Interspeech 2013), ISCA, Lyon, France, pp. 1877–1878.

Brandstätter, J., Moosmüller, S., 2015. Neutralisierung der hohen ungerundeten Vokale in der Wiener Standardsprache – A sound change in progress? In: Glauniger, M., Lenz, A. (Eds.), Standarddeutsch in Österreich – Theoretische und empirische Ansätze. Vandenhoeck & Ruprecht (Wiener Arbeiten zur Linguistik), pp. 183–203.

Brandstätter, J., Kaseß, C., Moosmüller, S., 2015. Quality and quantity in high vowels in Standard Austrian German. In: Leeman, A., Kolly, M., Dellwo, V., Schmid, S. (Eds.), Trends in Phonetics and Phonology in German Speaking Europe. Peter Lang, Frankfurt/Main, pp. 79–92.

Dressler, W.U., Wodak, R., 1982. Sociophonological methods in the study of sociolinguistic variation in Viennese German. Lang. Soc. 11, 339–370.

Fant, G., 1965. Formants and cavities. In: Proceedings of the 5th International Congress on Phonetic Sciences, Münster, Germany, pp. 120–141.

Gillick, L., Cox, S., 1989. Some statistical issues in the comparison of speech recognition algorithms. In: Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP'89), Glasgow, UK, pp. 532–535.

Goel, N., Thomas, S., Agarwal, M., Akyazi, P., Burget, L., Feng, K., Ghoshal, A., Glembek, O., Karafiat, M., Povey, D., Rastrow, A., Rose, R.C., Schwarz, P., 2010. Approaches to automatic lexicon learning with limited training examples. In: Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Dallas, Texas, USA, pp. 5094–5097.

Hershey, J., Olsen, P., Rennie, S., 2007. Variational Kullback–Leibler divergence for Hidden Markov Models. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, 2007. ASRU, Kyoto, Japan, pp. 323–328.

Hornung, M., Roitinger, F., 2000. Die österreichischen Mundarten. Eine Einführung. öbv&hpt, Wien.

Kawahara, H., Masuda-Katsuse, I., Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Commun. 27, 187–207.

Kranzmayer, E., 1956. Historische Lautgeographie des gesamtbairischen Dialektraumes. Österreichische Akademie der Wissenschaften, Wien.

Labov, W., 1972. Sociolinguistic Patterns. Blackwell, Oxford.

Lecumberri, M.L.G., Barra-Chicote, R., Ramón, R.P., Yamagishi, J., Cooke, M., 2014. Generating segmental foreign accent. In: Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 1302–1306.

Loots, L., Niesler, T., 2011. Automatic conversion between pronunciations of different English accents. Speech Commun. 53 (1), 75–84.

Moosmüller, S., 1991. Hochsprache und Dialekt in Österreich. Soziophonologische Untersuchungen zu ihrer Abgrenzung in Wien, Graz, Salzburg und Innsbruck. Böhlau, Wien.

Moosmüller, S., 2015. Methodisches zur Bestimmung der Standardaussprache in Österreich. In: Glauniger, M., Lenz, A. (Eds.), Standarddeutsch in Österreich – Theoretische und empirische Ansätze. Vandenhoeck & Ruprecht (Wiener Arbeiten zur Linguistik), Vienna, Austria, pp. 163–182.

Moosmüller, S., Brandstätter, J., 2014. Phonotactic information in the temporal organization of Standard Austrian German and the Viennese dialect. Lang. Sci. 46, 84–95.

Moosmüller, S., Vollmann, R., 2001. Natürliches Driften im Lautwandel: Die Monophthongierung im österreichischen Deutsch. Z. Sprachwissenschaft (20/I), 42–65.

Moosmüller, S., Schmid, C., Brandstätter, J., 2015. Standard Austrian German. J. Int. Phonet. Assoc. 45/3.

Müller, M., 2007. Dynamic time warping. Information Retrieval for Music and Motion. Springer, pp. 69–84.

Neubarth, F., Haddow, B., Hernandez-Huerta, A., Trost, H., 2013. A hybrid approach to statistical machine translation between standard and dialectal varieties. In: Proceedings of the 6th Language & Technology Conference (LTC'13), Poznan, Poland, pp. 414–418.

Nguyen, T.T.T., d'Alessandro, C., Rilliard, A., Tran, D.D., 2013. HMM-based TTS for Hanoi Vietnamese: issues in design and evaluation. In: Proceedings of the 14th Conference of the International Speech Communication Association (Interspeech 2013), Lyon, France, pp. 2311–2315.

Picart, B., Drugman, T., Dutoit, T., 2011. Continuous control of the degree of articulation in HMM-based speech synthesis. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH), Florence, Italy, pp. 1797–1800.

Pucher, M., Schabus, D., Yamagishi, Y., Neubarth, F., Strom, V., 2010. Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis. Speech Commun. 52, 164–179.

Pucher, M., Kerschhofer-Puhalo, N., Schabus, D., Moosmnller, S., Hofer, G., 2012. Language resources for the adaptive speech synthesis of dialects. In: Proceedings of the 7th Congress of the International Society for Dialectology and Geolinguistics (SIDG), Vienna, Austria, pp. 1–2.

Rabiner, L., Rosenberg, A.E., Levinson, S.E., 1978. Considerations in dynamic time warping algorithms for discrete word recognition. IEEE Trans. Acoust., Speech, Signal Process. 26, 575–582.

Richmond, K., Clark, R.A.J., Fitt, S., 2010. On generating Combilex pronunciations via morphological analysis. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH), Makuhari, Japan, pp. 1974–1977.

Russell, M., DeMarco, A., Veaux, C., Najafian, M., 2013. What's happening in accents & dialects? A review of the state of the art. In: UKSpeech Conference, Cambridge, 17/18 September 2013, Cambridge, UK.

Schabus, D., Pucher, M., Hofer, G., 2014. Joint audiovisual Hidden semi-Markov model-based speech synthesis. IEEE J. Sel. Topics Signal Process. 8 (2), 336–347.

Scheutz, H., 1999. Umgangssprache als Ergebnis von Konvergenz- und Divergenzprozessen zwischen Dialekt und Standardsprache. In: Stehl, T. (Ed.), Dialektgenerationen, Dialektfunktionen, Sprachwandel. Gunter Narr Verlag, Tübingen, Germany, pp. 105–131.

Soukup, B., Moosmüller, S., 2011. Standard language in Austria. In: Kristiansen, T., Coupland, N. (Eds.), Standard Languages and Language Standards in a Changing Europe. Novus Press, Oslo, Norway, pp. 39–46.

Springer, M., Thompson, W., 1970. The distribution of products of Beta, Gamma and Gaussian random variables. SIAM J. Appl. Math. 18 (4), 721–737.

Stan, A., Watts, O., Mamiya, Y., Giurgiu, M., Clark, R., Yamagishi, J., King, S., 2013. TUNDRA: a multilingual corpus of found data for TTS research created with light supervision. In: Proceedings of the 14th Conference of the International Speech Communication Association (Interspeech 2013), Lyon, France, pp. 2331–2335.

Tachibana, M., Yamagishi, J., Masuko, T., Kobayashi, T., 2005. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. IEICE Trans. Inform. Syst. E88-D (11), 2484–2491.

Tokuda, K., Kobayashi, T., Imai, S., 1995. Speech parameter generation from HMM using dynamic features. In: Proceedings of the 1995 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, Detroit, MI, USA, pp. 660–663.

Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T., 1999. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In: Proceedings of the 1999 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, Phoenix, AZ, USA, pp. 229–232.

Toman, M., Pucher, M., Schabus, D., 2013a. Cross-variety speaker transformation in HSMM-based speech synthesis. In: Proceedings of the 8th ISCA Workshop on Speech Synthesis (SSW), Barcelona, Spain, pp. 77–81.

Toman, M., Pucher, M., Schabus, D., 2013b. Multi-variety adaptive acoustic modeling in HSMM-based speech synthesis. In: Proceedings of the 8th ISCA Workshop on Speech Synthesis (SSW), Barcelona, Spain, pp. 83–87.

Valentini-Botinhao, C., Toman, M., Pucher, M., Schabus, D., Yamagishi, J., 2014. Intelligibility analysis of fast synthesized speech. In: Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH), Singapore, pp. 2922–2926.

Watts, O., Stan, A., Clark, R., Mamiya, Y., Giurgiu, M., Yamagishi, J., King, S., 2013. Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis. In: Proceedings of the 8th ISCA Workshop on Speech Synthesis (SSW). ISCA, pp. 101–106.

Wu, Y.-J., Nankaku, Y., Tokuda, K., 2009. State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH), Brighton, United Kingdom, pp. 528–531.

Yamagishi, J., Kobayashi, T., 2007. Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. IEICE Trans. Inform. Syst. E90-D (2), 533–543.

Yamagishi, J., Watts, O., 2010. The CSTR/EMIME HTS system for Blizzard challenge 2010. In: Proceedings of the Blizzard Challenge Workshop, Kansai Science City, Japan, pp. 1–6.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In: Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH), Budapest, Hungary, pp. 2374–2350.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speaker interpolation for HMM-based speech synthesis system. J. Acoust. Sci. Jpn. (E) 21 (4), 199–206.

Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Commun. 51 (11), 1039–1064.

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 74 (2015) 52–64

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Intelligibility of time-compressed synthetic speech: Compression method and speaking style

Cassia Valentini-Botinhao [a,*], Markus Toman [b], Michael Pucher [b], Dietmar Schabus [b], Junichi Yamagishi [a,c]

[a] The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
[b] Telecommunications Research Center Vienna (FTW), Austria
[c] National Institute of Informatics, Japan

Received 31 March 2015; received in revised form 17 August 2015; accepted 3 September 2015
Available online 25 September 2015

Abstract

We present a series of intelligibility experiments performed on natural and synthetic speech time-compressed at a range of rates and analyze the effect of speech corpus and compression method on the intelligibility scores of sighted and blind individuals. Particularly we are interested in comparing linear and non-linear compression methods applied to normal and fast speech of different speakers. We recorded English and German language voice talents reading prompts at a normal and a fast rate. To create synthetic voices we trained a statistical parametric speech synthesis system based on the normal and the fast data of each speaker. We compared three compression methods: scaling the variance of the state duration model, interpolating the duration models of the fast and the normal voices, and applying a linear compression method to the generated speech waveform. Word recognition results for the English voices show that generating speech at a normal speaking rate and then applying linear compression resulted in the most intelligible speech at all tested rates. A similar result was found when evaluating the intelligibility of the natural speech corpus. For the German voices, interpolation was found to be better at moderate speaking rates but the linear method was again more successful at very high rates, particularly when applied to the fast data. Phonemic level annotation of the normal and fast databases showed that the German speaker was able to reproduce speech at a fast rate with fewer deletion and substitution errors compared to the English speaker, supporting the intelligibility benefits observed when compressing his fast speech. This shows that the use of fast speech data to create faster synthetic voices does not necessarily lead to more intelligible voices as results are highly dependent on how successful the speaker was at speaking fast while maintaining intelligibility. Linear compression applied to normal rate speech can more reliably provide higher intelligibility, particularly at ultra fast rates.

Keywords: Fast speech; Time-compression; HMM-based speech synthesis; Blind individuals

1. Introduction

Blind individuals are capable of understanding speech reproduced at considerably high speaking rates. Comprehension is achieved even at a rate of 22 syllables per second (SPS), which is more than five times faster than the normal rate (Moos and Trouvain, 2007). Blind individuals often prefer to set their reading devices to reproduce speech at higher rates, changing the rate depending on how fast they would like to scan for information. For such applications, maintaining intelligibility, even if changes are detrimental to the naturalness, is essential. To be suitable for such specialized usage text-to-speech (TTS) systems should aim to provide synthetic speech that can remain intelligible at a range of rates.

* Corresponding author.
E-mail address: cvbotinh@inf.ed.ac.uk (C. Valentini-Botinhao).

http://dx.doi.org/10.1016/j.specom.2015.09.002
0167-6393/

There are four major types of speech synthesizers: formant, diphone, unit selection and statistical parametric synthesizer. The statistical parametric synthesizers have been found to be less intelligible at high speaking rates when compared to unit selection and a formant-based synthesizer (Syrdal et al., 2012). We believe however that this result could have been skewed by the compression method adopted by each synthesizer. In this work we focus on improving intelligibility of statistical parametric synthesizers as we believe they can provide additional advantages to the task. Parametric systems can generate intelligible speech from small databases, producing synthetic voices that match a certain speaker or accent without the need for long and high quality recordings. This can be particularly useful for blind children at a learning stage. It has been shown that familiarity with a speaker has a positive impact on the intelligibility of speech (Nygaard and Pisoni, 1998). This technology could provide children with screen readers and tutoring tools that use voices built from familiar persons like their teacher's, friends or even their own voice. If the same familiarity benefit also takes place with synthetic speech the use of such voices could improve their engagement. Recently Pucher et al. (2015) observed that blind children were more engaged when playing audio games made with synthetic voices built from recordings of people they knew. This was observed even though those speakers were not voice talents and there was less recording material.

The conventional acoustic model used in statistical parametric TTS is the hidden semi Markov model (HSMM) (Zen et al., 2007). HSMM-based synthesizers model duration by using explicit state duration distributions, usually a Gaussian distribution. To create speaking rates that are faster than the ones observed in the training data, the state-level duration can be increased by a factor which is proportional to the variance of the state duration model, following the duration elasticity of each segment. This method however has been shown not to produce particularly intelligible voices (Pucher et al., 2010; Syrdal et al., 2012). Pucher et al. (2010) has shown, for a particular speech database, that building a voice with fast speech data and interpolating the duration models of that and a normal voice creates speech that is more intelligible.

In this work, we are interested in examining two aspects of providing fast synthetic voices: the corpus used to train the voice and the method used for time compression. We believe that the performance of TTS systems can be improved by the adoption of an appropriate database and a more efficient compression method. For these reasons in the following paragraphs we describe aspects of speech production that can contribute to intelligibility at high rates and the different types of algorithms that have been used to compress natural and synthetic speech.

It is possible to create fast synthetic speech by training a voice with speech produced at fast rates. When producing fast speech, vowels are compressed more than consonants (Gay, 1978) and both word-level (Janse et al., 2003) and sentence-level (Port, 1981) stressed syllables are compressed less than unstressed ones. Yet another important aspect of fast speech is the significant reduction of pauses. It is claimed that reducing pauses is possibly the strongest acoustic change when speaking faster (Goldman-Eisler, 1968), most probably due to the limitations of how much speakers can speed up their articulation rate (Greisbach, 1992). These observed changes can be the result of an attempt to preserve the aspects of speech that carry more information, i.e. stressed segments. Silence removal is beneficial to a certain extent (Maxemchuk, 1980) as the presence of pauses helps maintain semantic and syntactic cues important for comprehension (Arons, 1992; Arons, 1994). The presence of pauses has actually been shown to contribute to intelligibility in certain scenarios (Sanderman and Collier, 1997), which would make heavily reducing pauses seem like a bad strategy.

During fast speech production, less acoustic or articulatory targets are reached, which makes fast speech sound mumbled and less intelligible. Foulke and Sticht (1969) observed that comprehension decreases slowly up to a word rate of 275 wpm (English) but more rapidly beyond that point. It has been shown that fast speech (around 1.56 times faster than normal speech – 6.7 and 10.5 SPS for normal and fast) is harder to process, in terms of reaction time, and also that it is preferred less than linearly compressed speech (Janse et al., 2003; Janse, 2004). Following the literature, the linearity here refers to the fact that the compression rate is the same across the sentence, i.e. vowels, consonants, silence and speech will be compressed at the same rate. Linearly compressed speech was found to be more intelligible and preferred over a nonlinearly compressed version of speech where fast speech prosodic patterns were mimicked (Janse et al., 2003) at a high speaking rate (2.85 times).

Janse (2004) claims that possibly the only nonlinear aspect of natural fast speech duration changes that can improve intelligibility at high speaking rates is the removal of pauses but only when rates are relatively high, as was shown in results obtained by another nonlinear compression method, the MACH1 algorithm (Covell et al., 1998). This method is based on the acoustics of fast speech with the addition of compressed pauses. At ultra fast speaking rates (2.5 and 4.1 times) MACH1 improves comprehension and is preferable to linearly compressed speech but no advantage was found at the fast speech speaking rate (1.4 times) (He and Gupta, 2001). Similarly Moers et al. (2010a) reported with a different fast speech corpus (2.0 times – SPS of 4 and 8 for normal and fast) that linearly compressed normal (plain) speech is more intelligible than fast speech but for an ultra fast speaking rate (rate was not reported) a linearly compressed version of the fast corpus was more intelligible than the plain counterpart and chosen to be more natural. Although there is a clear relation between the speed of articulation and the benefit of fast speech over linearly compressed speech, there is also something to be said about the different strategies that each

person adopts when speaking fast. Due to speaker variability, benefits seen at one rate for one speaker might not necessarily transpose to fast recordings of another speaker.

Fast natural speech was found to be more intelligible than fast speech generated by a formant-based synthesizer and the intelligibility gap grows with the speaking rate (Lebeter and Saunders, 2010). Moers et al. (2010b) reported that the use of units made of phones prone to heavy coarticulation when building a unit selection synthesizer with fast data creates speech that is more preferable in terms of intelligibility and naturalness than using phone units. The authors claim this counter balances the coarticulation reduction phenomena by adding contextual information. More recently Syrdal et al. (2012) evaluated the intelligibility of a wider range of synthesizers: formant, diphone, unit selection and HMM-based. It was found that the unit selection systems were more intelligible across speech rates. In this evaluation, however, the evaluated synthesizers were based on different speakers and the compression methods adopted by each system were not reported. Literature on fast synthesized speech also focuses on the effect on blind listeners as expert users of this technology. Intelligibility of formant and concatenative synthesizers was compared in Stent et al. (2011) for individuals with early onset blindness. It was found that recognition levels depend on the speaking rate and factors such as age, familiarity with the particular synthesizer and voice. In general, authors found that familiarity with the synthetic voice can alleviate the intelligibility drop that grows with rate, as a particular formant synthesizer voice outperformed other voices.

To improve duration control of HSMM-based systems for blind individuals (Pucher et al., 2010) proposed a model interpolation method that requires fast speech recordings. Pucher et al. (2010) found that interpolating between a model trained with normal and a model trained with fast speech data results in speech that is more intelligible and preferable, for both blind and non blind individuals. The most successful method in that study was one that applied interpolation between duration models only, using the normal speaking rate models of spectral and excitation related features. It is however not clear whether using recordings of fast speech is better than linearly compressing HSMM-generated synthetic speech and whether their results might change for a different speaker.

As previously mentioned in this work, we are interested in analyzing two aspects of producing fast synthesized speech. First the compression method; which is more effective: a nonlinear manipulation of speech duration or a linear compression method? Second the corpus used to train synthesis models, i.e., is it helpful to use fast speech recordings? To answer these questions we evaluate intelligibility of a fast and a normal female Scottish voice and an Austrian male voice, natural and synthetic, compressed using two nonlinear and one linear method and presented to listeners at different rates. Results have been partially published in Valentini-Botinhao et al. (2014). Here we extend them by presenting intelligibility scores of the German natural speech with both sighted and blind individuals as well as an analysis of the speech data derived from a manual phonetic annotation.

It is important to note here that it is not our goal to perform a cross lingual analysis of the effect of the compression methods and speaking styles. We are interested in identifying time compression methods that work best for synthesized speech. To train a high quality speaker dependent synthetic voice at least two hours of reasonable quality recordings (high sampling rate) are required, preferably from a voice talent. Although there are databases available that match these requirements, none, as far as we are aware, contain fast speech recordings as well. For those reasons we choose to work with a database of two different languages. The effect of this choice is further discussed later on.

This paper is organized as follows: Section 2 explains how speech is generated from a HSMM-based speech synthesizer, Section 3 describes the methods used to create synthetic speech at fast rates, Section 4 introduces the speech database and its phonetic annotation. Section 5 presents details on how the synthetic voices were trained using this data. Section 6 shows the design and results of intelligibility listening experiments, followed by a discussion in Section 7 and conclusions in Section 8.

2. HMM-based speech synthesis

Without explicit duration modeling, the state duration of an HMM would be given by the distribution of the transition probabilities which in turn give an exponential decaying distribution. As this is not a good model to generate natural sounding phone durations, explicit duration modeling in the form of the semi-Markov structure (Zen et al., 2007) was proposed. Under this framework, the duration of a particular state is modeled by a Gaussian distribution.

At generation time the text to be synthesized is converted to a sequence of linguistic specifications that drives the selection of context dependent HMM models. Given this concatenated sequence of HMMs, the utterance HMM λ is constructed. The most likely observation sequence is given by the maximization of the likelihood function. Tokuda et al. (2000) showed that a closed form solution can be found if we consider just the most likely state sequence:

$$\boldsymbol{q}_{max} = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q}|\lambda) \tag{1}$$

where $\boldsymbol{q} = \{q_1, q_2, \ldots, q_T\}$ and T the number of states.

Assuming the HMM state transition goes from left-to-right without skipping states, it is possible to find the state sequence from the model by using the state duration probability, which in the HSMM paradigm is explicitly modeled by the distribution:

$$\boldsymbol{q}_{max} = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q}|\lambda) \quad (2)$$

$$= \arg\max_{\boldsymbol{q}} \prod_{k=1}^{K} p_k(d_k) \quad (3)$$

where $p_k(d_k)$ is the probability of duration d_k in state k, i.e. the probability that a segment of d_k duration is emitted from state k; K is the number of states visited during the duration of T – given by the model specification. The total duration has to be achieved so $\sum_{k=1}^{K} d_k = T$. For the Gaussian distribution, the state durations that maximize Eq. (3) is given by (Yoshimura et al., 1998):

$$d_k = \mu_k + \rho\,\sigma_k^2 \quad (4)$$

$$\rho = \left(T - \sum_{k=1}^{K} \mu_k\right) \Big/ \sum_{k=1}^{K} \sigma_k^2 \quad (5)$$

where μ_k and σ_k are the mean and the variance of the duration distribution of state k and ρ is a parameter that can be controlled by a desired total duration as we will soon show. From the values of d_k, we know how many frames are emitted by each state and therefore the state sequence. When synthesizing a sentence it is possible to set a desired total duration T. From the equations above, we are able to calculate the state duration d_k, however, rounding errors in the process of approximating a real value (time) to a integer value (number of states) means that the generated sentence will not necessarily have exactly duration T.

3. Time compression methods

In this section, we describe some of the methods that can be used to create synthetic speech at fast rates, referred to here as time compression methods. The first two methods manipulate speech at the acoustic model level by modifying the state duration model parameters (mean and/or variance) while the third is applied to the synthesized speech waveform. The first two methods are considered to be non-linear as each state is compressed at a different rate, as opposed to the third method, which is a linear method that compresses the waveform relatively uniformly across time and different speech units.

3.1. Variance scaling

Variance scaling is proposed as the standard method for duration control in HSMM-based synthesis (Yoshimura et al., 1998). At generation time duration of state i is computed as Eq. (4). When $\rho = 0$ the duration is set to the mean state duration. Under this setting the synthesized voice should have the normal speaking rate of the training data. Turning $\rho > 0$ makes the synthetic speech slower and $\rho < 0$ makes it faster. The scaling factor is fixed across all states. State duration control is then proportional only to the variance: states whose duration model variance is higher will be compressed more. With this method we can potentially capture certain non-linearities between normal and fast speech durations seen in the training data. In the same way as in fast speech, vowels are more compressed than consonants, this method should also compress such units more as the duration model variance of states referring to vowels is higher (Pucher et al., 2010).

3.2. Model interpolation and extrapolation

In previous work with fast synthetic speech, Pucher et al. (2010) showed that model interpolation (Yoshimura et al., 2000; Tachibana et al., 2005) can outperform the variance scaling method in terms of intelligibility and listener preference. Given two voice models of the same speaker trained with speech recorded at normal and fast speaking rates, the most successful method in that study was one that applied interpolation between duration models only, using the normal speaking rate models of cepstral, fundamental frequency and aperiodicity features. The interpolated duration d_i for state i is calculated as:

$$d_i = (1-\alpha)\,\mu_i^n + \alpha\,\mu_i^f \quad (6)$$

where μ_i^n and μ_i^f denote the mean duration of state i in the normal and fast duration model and α is the interpolation ratio to control the speaking rate. We can generate speaking rates beyond the rate of the fast model by extrapolating ($\alpha > 1$). Interpolation itself is a linear method at a state level as the duration of each state is modified by a linear term. However because the duration change is different for different states, i.e. different speech segments, we classify it here as a non linear time compression method.

For the experiments in the present paper, we have implemented an additional constraint in this method. It is theoretically possible that for a given state of a given phone, the mean duration μ_i^f from the fast model may be actually longer than the mean duration μ_i^n of the normal model, causing the speech segments generated for this state to become *slower* with growing α. If this is the case, we do not interpolate or extrapolate, but apply a linear factor β to μ_i^n, where β reflects the overall mean speaking rate difference between the normal and the fast voice models ($\beta = 1/1.55$ in our experiments).

3.3. WSOLA

The waveform similarity overlap and add (WSOLA) method proposed in Verhelst and Roelands (1993) was chosen here to illustrate the effect of a linear compression. The method provides high enough quality while being computationally efficient and robust (Verhelst and Roelands, 1993). In WSOLA, speech frames to be overlapped are first cross-correlated to provide an appropriate time shift that ensures frames are added coherently, inspired by the idea that modified speech should maintain maximum local similarity to the original signal. We say here that the compression is relatively linear across time as the method allows a tolerance region within which a window can be placed in

order to allow for a maximum similarity to be reached. The similarity measured can be for instance the cross correlation or the cross AMDF (average magnitude difference function) between the downsampled sequence underlying a particular segment and that segment input sequence.

4. Database description and annotation

We recorded a Scottish female voice talent reading 4600 sentences at a normal speed and 800 sentences (contained in the normal read prompts) at a fast speed. The instruction was to speak as fast as possible while maintaining intelligibility.

Using a similar setup we recorded an Austrian German voice talent reading 4387 sentences at a normal and 198 sentences at a fast speaking rate.

We performed a manual annotation of the natural speech data at a phonemic level, i.e. determining which phonemes were pronounced and where their boundaries lie. This section presents details of the annotation and its results.

4.1. Annotation procedure

We annotated phone and syllable level information, i.e. time stamps and phone identification, of recordings of each speaker reading the same sentence material in two different speaking styles: normal and fast.

To choose the sentences to be annotated we drew 40 sentences out of the 144 sentences used for the listening test with natural speech. As it is easier for the annotator to start with at least a segmentation rather than doing everything from scratch we provided initial segmentations by retrieving duration generated by the models built for that speaker and those sentences. With the label files generated after forced alignment (performed using the TTS models trained with the corresponding data) and the wavefiles, the annotators used Praat (Boersma and Weenink, 2014) to perform the task. Figs. 1 and 2 show a Praat window for one sentence in the English and German corpus. Here we can see the phone, syllable and orthographic levels and the boundaries set by the annotators.

As we are interested in comparing differences across styles it was important to be able to annotate the changes observed by our annotators. In general, transcription was fairly broad i.e. phonemic, however, particularly for the fast speech, it was noted that a slightly narrower transcription would be desirable, as the speaker employed certain strategies in order to speed up, such as not achieving closure for plosives (indicated in Figs. 1 and 2 by the symbol *), and occasionally not releasing plosives (indicated by the symbol ~). Because it is not necessary to have a fully narrow transcription at the phonetic level, but we do want to know about such cases where plosives are not fully realized, symbols were added to describe unreleased plosives and plosives without complete closure.

Syllable boundaries were generally kept faithful to the canonical syllabification as provided by the pronunciation dictionary. The syllable boundaries were not changed unless there were deletions which then changed the syllabification (e.g. something deleted a syllable boundary). Although one approach could have been to attempt to model resyllabification by the speaker – particularly in fast speech, where for example the coda of one syllable becomes the onset of the next – this was not attempted due to the arising ambiguity around such cases, and the need for consistency. Syllable boundaries are therefore always within-word or at word boundaries and do not span word boundaries.

The English data was annotated by an experienced annotator who is a native English speaker. The German data was annotated by a German native speaker and cross checked by the English annotator for consistency across languages. It was noted by the annotators that phoneme boundary decisions were easier to make in the German fast data than in the English fast data, as the articulation changes were more pronounced and therefore visually present in spectrograms, as seen in the example of Figs. 1 and 2. In this example the syllable per second and increase in speaking rate relative to normal speech is comparable across languages, yet the spectrogram of the German sentence is more contrastive.

4.2. Annotation results

We present results for both languages (speakers) in Table 1. The total number of phones annotated in the normal data is presented in the first column. The other

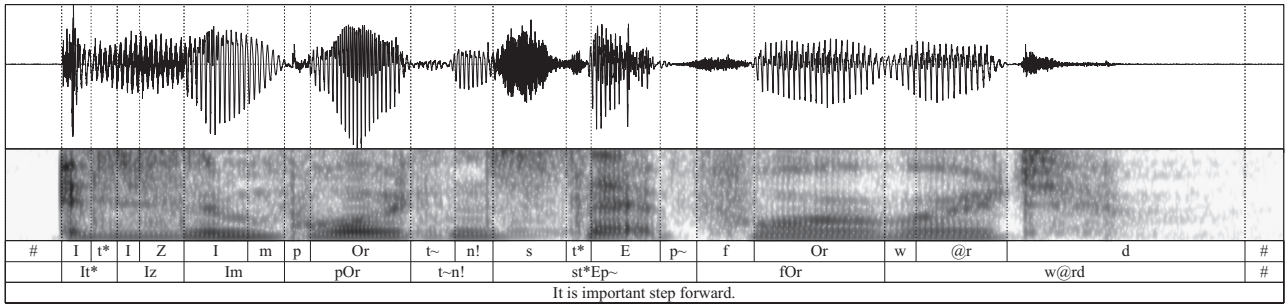


Fig. 1. English fast sentence in Praat, phone symbols are from the Combilex lexicon (Richmond et al., 2010).

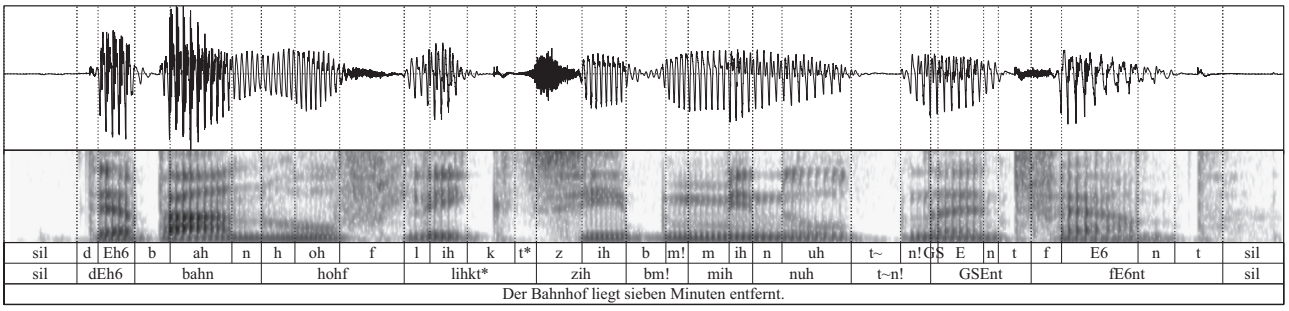


Fig. 2. German fast sentence in Praat, phone symbols are from the Austrian German lexicon (Neubarth et al., 2008).

columns present hits (phone correct assignment), deletions, substitutions and insertions of phones annotated in the fast speech compared to phones annotated in the normal speech for both English and German. Numbers in parenthesis refer to unique counts, i.e. how many different instances have occurred. The last column refers to the phone error rate calculated for the fast speech annotation having the normal speech annotation as the reference. This should give us a reasonable indication of how intelligible the fast data is. For both German and English we annotated 40 sentences but some sentences of the German data were particularly long, so more phones were annotated for that language.

Figs. 3 and 4 present the pattern of compression rate and absolute decrease in duration (in ms) across different phonetic units. To calculate this we took the average across phone occurrences within each phonetic class, the standard deviation refers to the deviation across phone appearances. Only phones that occurred more than five times in the annotation were considered to remove any outliers. Deletions were not counted here, but pointed out separately in the table mentioned previously. The phonetic classes considered were: vowels, nasals, stops, approximants and fricatives.

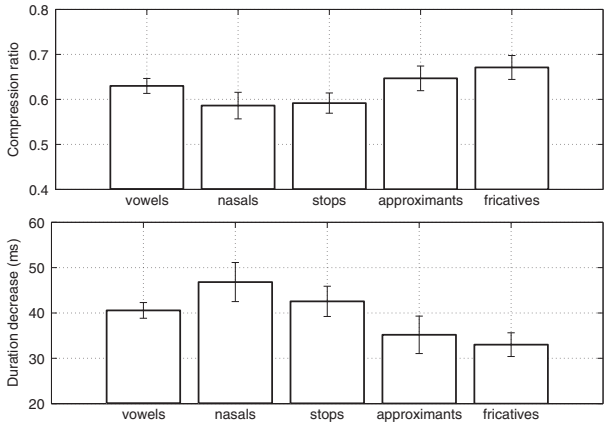


Fig. 3. English annotation results: compression ratio (top) and absolute duration decrease in ms (bottom).

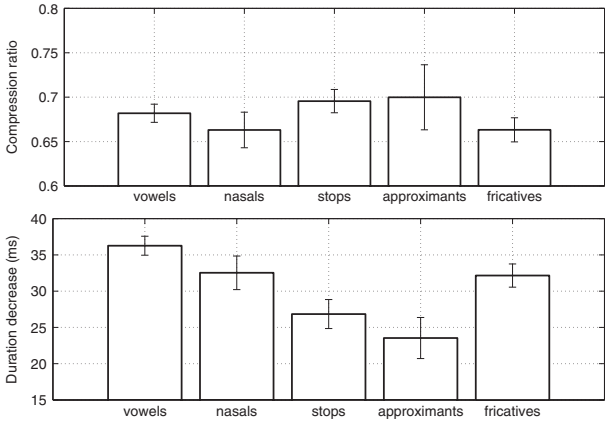


Fig. 4. German annotation results: compression ratio (top) and absolute duration decrease in ms (bottom).

4.2.1. English

For the English data, as seen in Table 1, 857 phones were annotated from 40 sentences of normal speech, 56 of which were unique. Contrasting the normal with the fast annotation we found that 84% were present in the fast data while the rest was either deleted or replaced. The most common substitutions were plosives being replaced by unreleased plosives (35 substitutions) or incomplete closure in plosives (18), vowel reductions to schwa (14) and [t] becoming a glottal stop (5). The most common deletions were pause removals (9 times), [t] (7), schwa (6) and, [l]

Table 1
Annotation results: number of phones annotated in the normal data, number of phones hits, deletions, substitutions and insertions in the fast data. Numbers in parenthesis refer to unique counts. PER refers to phone error rate of the fast speech transcription compared to the normal speech.

	Phones	Hits	Deletions	Substitutions	Insertions	PER
English	857(56)	717/84%	45(16)/5%	95(39)/11%	0	16.34%
German	1480(67)	1359/91%	62(18)/4%	55(36)/4%	4(4)	8.20%

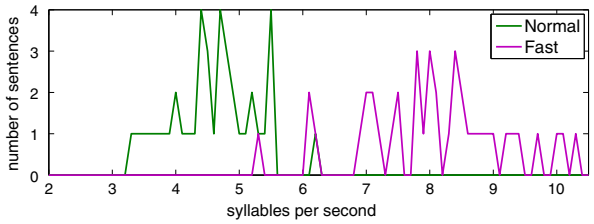


Fig. 5. English: syllables per second distribution calculated from annotation. Average values for Normal and Fast: 4.7 and 8.1.

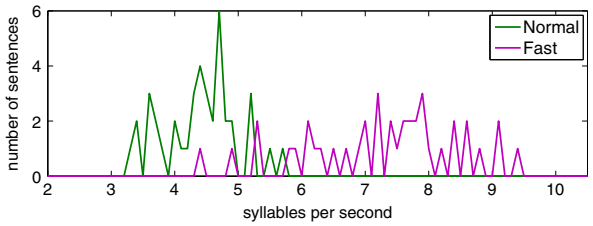


Fig. 6. English: syllables per second distribution calculated from labels. Average values for Normal and Fast: 4.5 and 7.3.

and [h] (4 times each). The phone error rate of the fast speech data is of 16.34%.

To understand the durational changes made by the speaker we present in Fig. 3 the compression ratio and duration decrease per phonetic class. We can see that this speaker compressed nasals and stops the most and fricatives and approximants the least. The absolute decrease in duration for stops and vowels was similar, vowels were however expected to shorten more according to the literature on fast speech of other speakers Gay (1978).

Fig. 5 presents the distribution of SPS across the annotated sentences for normal and fast data. The average words per minute (WPM) values are 215.0 (217.6 discarding pauses) for the normal case and 371.8 for the fast case[1]. The average SPS for the normal case is 4.7 (4.8 discarding pauses) and fast 8.1. That means that the speaking rate with pause was found to be 1.76 (discarding pause 1.78) which is higher than we calculated from the labels that trained the models. Forced alignment using an HSMM model trained with the data gives slightly smaller values for the same 40 sentences, as can be seen in Fig. 6, on average 4.5 and 7.3. We found that in some cases the initial phone includes a large part of the preceding silence, which means that speech duration according to labels is longer than the manual annotation and this issue is more pronounced when automatically segmenting the fast speech.

4.2.2. German

For the German data 1480 phones were annotated from 40 sentences of normal speech, 67 of which were unique. Contrasting the normal with the fast annotation we found that 91% phones were present in the fast data while the rest were either deleted, substituted or inserted. The most common substitutions were incomplete closure in plosives (7 times), replacement of schwa by [e] (5 times), replacement of [eɒ] by [e:ɒ] (4 times), and vowel reductions to schwa (4). The most common deletions were pause removals (27 times), removal of glottal stop (7), and removal of [t] (6), [r] (4), [d] (3). Furthermore we had insertions of glottal stop and [g] (1 each), and two insertions for splitting up compound phones. The phone error rate of the fast speech data is 8.2%, which is much lower than for the English speaker. This also shows that the German speaker produces more consistent speech output when speaking fast.

Fig. 4 shows the compression ratio and duration decrease per phonetic class for the German speaker. We can see that this speaker compressed the most nasals and fricatives and the least stops and approximants. The absolute decrease in duration for stops and vowels was different, as expected according to the literature on fast speech of other speakers.

Fig. 7 presents the distribution of SPS across the annotated sentences for normal and fast data. The average WPM value for the normal case is 157.9 (164.0 discarding pauses) and 267.7 for the fast case. The average number of SPS for the normal case is 4.2 (4.3 discarding pauses) and fast 7.0. Forced alignment using an HSMM model trained with the German data gives similar SPS values, as can be seen in Fig. 8.

5. Synthetic voices

In this section, we present details on how we trained the synthetic voices. The synthetic voice described as the fast voice was built by adapting the duration model only to the fast speech data as Pucher et al. (2010) reported it results in more intelligible voices.

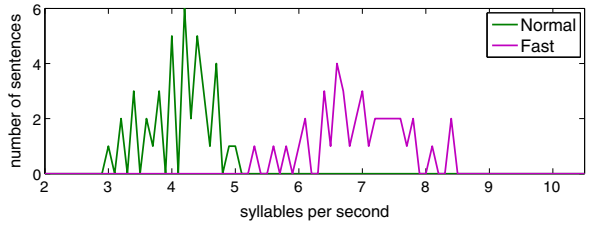


Fig. 7. German: syllables per second distribution calculated from annotation. Average values for Normal and Fast: 4.2 and 7.0.

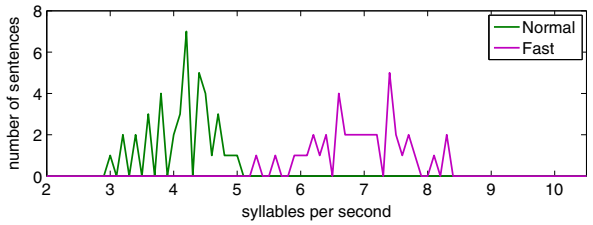


Fig. 8. German: syllables per second distribution calculated from labels. Average values for Normal and Fast: 4.2 and 7.0.

[1] Bearing in mind that (Foulke and Sticht, 1969) found comprehension decreases rapidly above 275 wpm.

5.1. English voices

To train the acoustic models, we extracted the following features from the natural speech sampled at 48 kHz: 59 Mel cepstral coefficients (Fukada et al., 1992), Mel scale fundamental frequency (F0) and 25 aperiodicity band energies extracted using STRAIGHT (Kawahara et al., 1999). We used a hidden semi-Markov model as the acoustic model. The observation vectors for the spectral and excitation parameters contained static, delta and delta-delta values. One stream was set for the spectrum, three for F0 and one for aperiodicity. The lexicon used by the front-end was the Combilex lexicon (Richmond et al., 2010). At the time of this work only the British Received Pronunciation accent variant of Combilex was available.

We trained two voices. What we refer to as model N, is a voice trained only with speech produced at the normal speaking rate. The voice duration model of model N was adapted (Yamagishi et al., 2009) using the 800 sentences of fast speech to create what is referred to as voice F.

To measure the speaking rate of each synthetic voice we calculated the rate of SPS and WPM for each sentence used in the evaluation by performing forced alignment on the data using the corresponding N and F models. On average the SPS values of the normal and the fast voice are 3.8 and 6.0 while the values for WPM are 206.7 and 320.9, respectively. Speech synthesized using the fast model is around 1.55 times faster, which agrees with the literature (Janse et al., 2003) on naturally produced fast speech (1.56 times faster). Fig. 9 shows the histogram of SPS across synthesized sentences for each voice.

5.2. German voices

The German recordings were sampled at 44.1 kHz and we extracted 39 Mel cepstral coefficients. Otherwise the procedure and parameters were the same as for English.

The average SPS values, calculated from forced alignment, for the normal and fast German synthetic voices are 4.5 and 7.0, which was quite similar to the results obtained for the natural speech database in Section 4, and the WPM values are 152.7 and 237.1. The German voice is thus considerably slower than the English voice, at both speaking rates. Interestingly, the fast model is also about 1.55 times faster than the normal model, i.e., the speed-up factor between the two English models and between the two German models is the same. Fig. 10 shows the SPS distribution for the two German models.

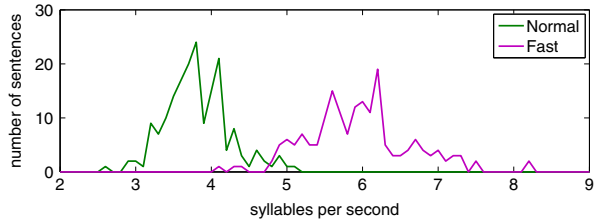


Fig. 9. English TTS voices: syllables per second distribution calculated from labels. Average values for Normal and Fast: 3.8 and 6.0.

6. Evaluation

We recruited native speakers of each language. Each participant had to type the words they could distinguish from the sentences that were played to them. If they could not distinguish any word in the sentence listeners should type 'x'. No listener heard the same sentence twice, and each condition was heard by the same number of listeners.

We conducted two listening experiments, one using natural speech and the other TTS. As the work that led to this article was supported by a project that involved a school for blind children in Austria it was straightforward to conduct listening tests with blind individuals in that country. Once such tests were performed we noted that the trend, i.e. the relative performance of each method, was quite similar to results with sighted participants. For this reason we decided not to perform an English language test with blind individuals.

Results are presented as percentage of word errors, calculated per listener as the percentage of words that were not found in the listener's transcription, misspellings taken into account, and then averaged across listeners. Following the same procedure adopted in the intelligibility evaluation described in (Cooke et al., 2013) we did not count insertions as errors. The bars refer to standard deviation of the error calculated across listeners' results.

6.1. Speaking rates

We evaluate intelligibility of time-compressed speech at four different speaking rates: 1.25, fast (the speed of fast speech), 2.0 and 3.0, where numbers refer to the speed increase with respect to the normal speech calculated sentence by sentence, remembering here that fast speech is around 1.55 times faster than normal speech. Rates were chosen to reflect conversational, fast, and two ultra fast speeds.

6.2. Methods

The acronyms for the corpus compression method combinations we evaluate are presented in Table 2.[2] Not all methods are evaluated at all speaking rates, for instance at rates smaller or equal to the fast rate W-F, V-F and I were not evaluated as that would mean slow down the speech signal. Also for the natural speech evaluation only the WSOLA algorithm was used as the other two methods cannot be applied directly to natural speech. To generate compressed samples using the variance and the interpolation methods it was necessary to progressively change the

[2] Speech samples used in the evaluation can be found at: http://wiki.inf.ed.ac.uk/CSTR/SalbProject.

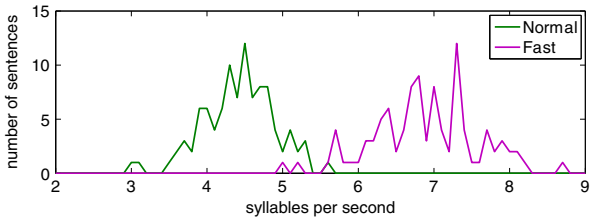


Fig. 10. German TTS voices: syllables per second distribution calculated from labels. Average values for Normal and Fast: 4.5 and 7.0.

scale factor to obtain the desired duration. The implementation of WSOLA used here was provided as support material in Rabiner and Schafer (2010). We used the AMDF as the similarity function and 5 ms as the size of the tolerance window as proposed in Verhelst and Roelands (1993).

6.3. English evaluation

We performed two listening experiments, one with natural speech and the other with the TTS voices. Each experiment was performed by 20 native English speakers without TTS expertise. Each participant transcribed 100 different sentences for each of the tested methods. The natural speech sentences were selected from news articles while for the TTS experiments sentences were chosen from the first few sets of the Harvard dataset (IEEE, 1969). We chose this material as opposed to semantically unpredictable sentences (Syrdal et al., 2012) because we are interested in testing real life scenarios, i.e. extremely high speaking rates. The SUS material is quite challenging and intentionally unrepresentative of realistic sentences, which would almost certainly result in intelligibility being lost at much slower speaking rates.

For the natural speech evaluation we compare two natural speech compressions: W-N and W-F, compression applied to the normal and the fast speech databases. All speaking rates were evaluated except the 1.25 rate which was considered too easy a task.

For the TTS evaluation, we compare the three different compression methods described in Section 3, although not all methods were evaluated for all speaking rates. For the English data it was not possible to generate speech at 3.0 rates with the variance (V) and interpolation method (I) due to the limitation of the minimum number of frames (no state skipping is allowed).

6.3.1. Results

Fig. 11 shows the percentage of word errors for each speaking rate obtained in the natural (blue) and TTS (red) experiments.

We can see that the TTS voices created using WSOLA are the most intelligible across all tested speaking rates and that this advantage grows with increasing speaking rate. At the 2x rate the TTS voice W-N results in less than 20% word errors while the word errors obtained by V-N and I are higher than 40%, i.e., errors doubled. Results for W-F and V-F are better, just around 30%.

Table 2
Methods evaluated.

W-N	WSOLA applied to normal speech
W-F	WSOLA applied to fast speech
V-N	Variance scaling applied to model N
V-F	Variance scaling applied to model F
I	Interpolation of models N and F

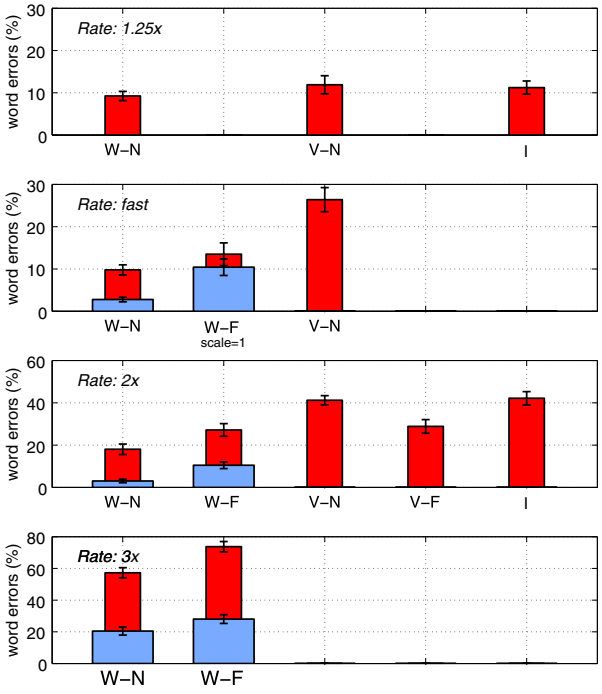


Fig. 11. English results: TTS (red) and natural speech (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Word errors are smaller when compressing speech synthesized from the normal model (W-N) as opposed to a fast model (W-F), as results for speaking rate 2x and 3x show. The error level for V-F is however smaller than V-N. At the fast speaking rate, we can see that the fast voice is less intelligible than the normal voice with linear compression applied.

The error scores for natural speech (in blue) are significantly lower than the ones obtained for the TTS voices, and the intelligibility gap grows with speaking rate. The increase in error seen for W-F when compared to W-N for TTS voices can also be observed for natural speech, pointing to the fact that the fast natural speech is also less intelligible than linearly compressed normal speech.

6.4. German evaluation

The participants in the listening test consisted of two groups: 16 blind or visually impaired participants, 15 of whom reported using TTS in their everyday life, and 16

sighted participants with no TTS expertise. Each participant transcribed 100 different sentences such that within a participant group, every combination of method and speaking rate was evaluated once. The sentences were selected from news articles and parliamentary speeches for the TTS experiment and from a similar corpus of recordings for the natural speech experiment.

6.4.1. Results

The results are shown in Fig. 12. As in the English experiments, the percentage of word errors for each speaking rate and for each method is shown for the natural (blue) and TTS (red) experiments.

Similar to the English results, WSOLA compression of speech synthesized from the normal model (W-N) is the best method overall. However, up to speaking rate 2x, both WSOLA of fast speech (W-F) and interpolation (I) yield results competitive to W-N. At the "fast" rate, where both W-F and I (and also V-F) are equivalent to simply the fast voice model, these methods even achieve significantly better results than W-N for the sighted listeners. At the "fast" and 2x rates, W-F and I perform significantly better than variance scaling of the normal model (V-N), confirming the results of Pucher et al. (2010). However, we see a very clear advantage of the WSOLA methods at the fastest rate 3x, where the error percentages of V-N, V-F and I are much higher, yielding a picture similar to the English results at 2x. There is no significant difference between W-N and W-F at the 2x and 3x rates.

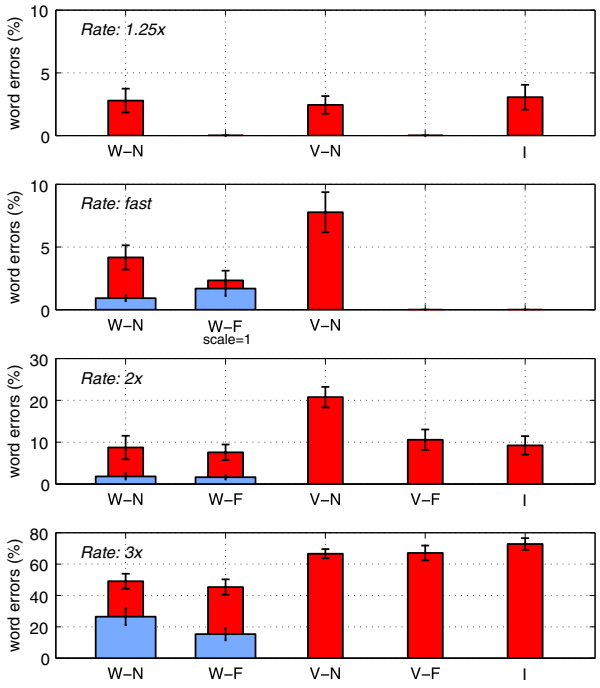


Fig. 12. German results: TTS (red) and natural speech (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

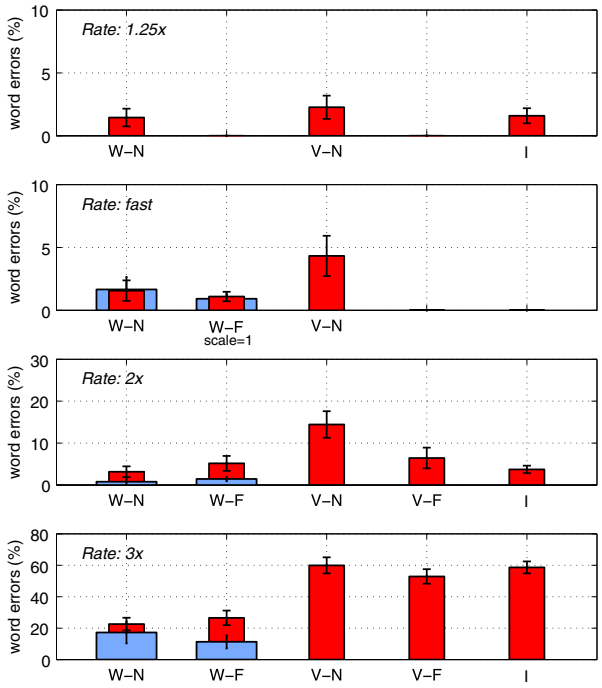


Fig. 13. German results with blind/visually impaired listeners: TTS (red) and natural speech (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

As observed in the English results, the natural speech error scores were significantly lower than the scores obtained with TTS and this gap grows with higher compression rates. At moderate rates, fast and 2.0, W-N and W-F results are comparable but at the highest rate compressing fast speech creates more intelligible stimuli, unlike the findings for the English data. This will be further investigated in the following section.

6.4.2. Results with blind/visually impaired

As shown in the literature (Stent et al., 2011; Trouvain, 2007; Pucher et al., 2010), blind listeners generally achieve lower word error percentages than sighted listeners. Fig. 13 presents the results with blind/visually impaired individuals. The overall picture concerning the different methods is very similar to the results of the experiments with sighted listeners, still the WSOLA methods perform best at higher speaking rates. The advantage of using a linear compression method seems even larger here, particularly at the highest speaking rate, where W-N and W-F errors were around 22%, a result much better than the one obtained by sighted participants, but V and I methods' results were still between 50% and 60% a result comparable to the one obtained by sighted listeners.

7. Discussion

As found in other studies of natural fast speech (Janse et al., 2003; Janse, 2004), our results using the English data also indicate that linear compression produces more

intelligible voices than nonlinear methods based on or directly derived from the acoustics of fast speech. English results show that there is no additional advantage to using recordings of fast speech to build a synthetic voice and it is possible to maintain intelligibility at higher speaking rates by applying a linear compression method such as WSOLA to the synthesized waveform. This is supported by results with the natural speech corpus, where we also found that fast natural speech is not as intelligible as linearly compressed normal speech.

Results for the German data tell a slightly different story. For German, we also see that linear compression is beneficial at very high speaking rates (3x) compared to interpolation and variance scaling. For lower speaking rates (2x), we find that interpolation is equally as good as linear compression. This indicates a potential use of a combined method of interpolation for fast speaking rates and linear compression for ultra-fast speaking rates. We hypothesized in Valentini-Botinhao et al. (2014) that different results were found for the German data due to the inherent higher intelligibility of the German fast speech, which can also be seen in the performance differences of linear compression of synthesized speech from fast models (W-F) which performs better for the German data. The results with natural speech confirmed this as we observed a significant improvement at the highest speaking rate when compressing the fast German data, a result that was not seen in the English data.

Concerning the performance of blind listeners we can confirm results presented in previous studies (Moos and Trouvain, 2007; Pucher et al., 2010), which show that blind listeners achieve lower word-error-rates than non-blind listeners. Moreover we observed that the WSOLA compression intelligibility gains were higher for blind users, most probably due to the fact that they are expert users of TTS systems that use similar types of compression. The intelligibility gap between WSOLA and variance scaling or interpolation was larger with blind individuals, particularly at higher rates, another compelling reason to use a linear method.

The annotation results showed that the German natural fast data is more intelligible than the English data in terms of how many phones were found to be present in the fast data and the number of deletions and substitutions. The phone error rate was found to be twice as high for the English data. The most common substitutions for both speakers were plosives becoming incomplete and vowels replaced by schwa. The deletions in German were dominated by pause removal, which could explain how the German speaker was able to reach more phonetic targets while still maintaining a relatively high speaking rate. The pattern of compression across different phonetic units showed that the German speaker presents a similar pattern as to the one often presented in the literature, i.e. highly compressed vowels and nasals, while the English speaker most compressed unit was found to be the stops. A larger amount of the corpus would have to be annotated to make further statements but at this point we were able to identify that there were in fact large differences between the compression strategies of each speaker and that one was better able to correctly produce speech at higher rate. We believe that the Austrian speaker was considerably more intelligible at high speaking rates because he is also a professional narrator and therefore more experienced at such tasks. To discard language effects we hope to extend this study using speaker adaptive voices built from limited amount of recordings of non professional speakers of the same language.

Considering results on both databases we hypothesize that methods that use recordings of fast speech such as adaptation or interpolation are perhaps only as intelligible as the fast data they use. Relying on having fast speech that is intelligible enough is challenging as this data is quite difficult to produce, particularly considering that both of our speakers are voice talents. Using more recordings of fast speech is also not helpful as more fast sentences were used for the English voices. Moreover it is not yet clear how to reach very high speaking rates with model interpolation and adaptation as these methods are limited by the fact that skipping states is not allowed during generation. The weak performance of the variance scaling method for fast speaking rates (2x, 3x) is in agreement with the poor results obtained by HMM-based voices in Syrdal et al. (2012).

One of the follow up questions is how to obtain fast speech recordings that are still clear enough to be useful to build synthesizers. In both recordings the speakers were instructed to speak as fast as they could while maintaining intelligible speech. The task was however to read sentence by sentence out loud. Some might argue that this is not the best way to collect fast speech data as it is difficult to keep a stable rate while reading and speaking at the same time, particularly when prompted to read sentence by sentence. One possible way to collect fast data would be to record conversational speech that is known to be produced at faster rates than read speech, the material usually used to train synthesizers. Training a TTS synthesiser with conversational speech is however still a challenging task for both the front-end, in terms of presenting alternative ways of pronouncing scripts, and the back-end, for both the vocoder and the acoustic model. A possible alternative to fast speech is thinking in the other direction, i.e., clear speech (Hazan and Baker, 2011). Clear speech is significantly more intelligible than conversational speech but to the expense of longer utterance duration. The question still remains whether speeding up clear speech can improve upon results with normal read speech.

Another question that rises from this study is how to further improve linear methods. We have seen large intelligibility advantages both for sighted (non expert TTS users) and visually impaired individuals when using a linear compression method. Could a method such as the MACH1 that makes use of silence removal further improve upon these results? Using WSOLA to manipulate synthetic speech requires an additional stage of analysis and

synthesis that could be avoided if the modification was performed at generation time. Currently most systems train speech with 5 ms displaced windows. If this number is made smaller the speaking rate would increase. We would expect however that such a method would cause more artefacts as overlap and add windows would be displaced without any regard to continuity, as opposed to the WSOLA method that defines where to place windows according to a similiarity measure (Verhelst and Roelands, 1993). A potentially more attractive modification would be to change the duration of each state linearly by ignoring the state variance and setting the final state duration to a proportion of the state duration model mean value according to the desired speaking rate. The parameters that describe the dynamics (the delta and delta–delta values of the observation vector) would however not be appropriate as they were learned using speech produced at normal rates. State transitions would then not be as rapid as they should be which could lead to badly articulated synthetic speech.

8. Conclusion

We presented results of intelligibility listening experiments with natural and synthetic speech of English and German voices reproduced at higher rates by a variety of methods. We showed that a linear compression method outperforms the variance scaling and interpolation methods, particularly for ultra-fast (3x) speaking rates. For fast speaking rates (2x) linear compression outperformed other methods for English while being as good as interpolation for German. In general we see that the usage of fast speech data in interpolation or linear compression is dependent on how intelligible the fast data is, where the rate of intelligibility can be determined by measuring the phone error between manually aligned annotations of normal and fast speaking rate utterances. Future work includes investigating the use of other sorts of speaking styles such as conversational speech and combination of linear and non-linear methods to further improve upon current results.

Acknowledgement

This work was supported by the Austrian Federal Ministry of Science and Research (BMWF) – Sparkling Science project *Sprachsynthese von Auditiven Lehrbüchern für Blinde SchülerInnen* (SALB). The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH is funded within the program COMET – Competence Centers for Excellent Technologies by BMVIT, BMWA, and the City of Vienna. The COMET program is managed by the FFG.

References

Arons, B., 1992. Techniques, perception, and applications of time-compressed speech. In: Proceedings 1992 Conference, American Voice I/O Society, pp. 169–177.

Arons, B., 1994. Interactively Skimming Recorded Speech. PhD thesis – MIT.

Boersma, P., Weenink, D. Praat: doing phonetics by computer [Computer program]. Version 5.3.75, retrieved 1 May 2014. <http://www.praat.org/>.

Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., Tang, Y., 2013. Evaluating the intelligibility benefit of speech modifications in known noise conditions. Speech Commun. 55, 572–585.

Covell, M., Withgott, M., Slaney, M., 1998. Mach1: nonuniform time-scale modification of speech. In: Proc. ICASSP, vol. 1. IEEE, Seattle, USA, pp. 349–352.

Foulke, W., Sticht, T.G., 1969. Review of research on the intelligibility and comprehension of accelerated speech. Psychol. Bull. 72, 50–62.

Fukada, T., Tokuda, K., Kobayashi, T., Imai, S., 1992. An adaptive algorithm for mel-cepstral analysis of speech. In: Proc. ICASSP, vol. 1, San Francisco, USA, pp. 137–140.

Gay, T., 1978. Effect of speaking rate on vowel formant movements. J. Acoust. Soc. Am. 63 (1), 223–230.

Goldman-Eisler, F., 1968. Psycholinguistics: Experiments in Spontaneous Speech. Academic Press, London.

Greisbach, R., 1992. Reading aloud at maximal speed. Speech Commun. 11 (4–5), 469–473, ISSN: 0167-6393.

Hazan, V., Baker, R., 2011. Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. J. Acoust. Soc. Am. 130 (4), 2139–2152.

He, L., Gupta, A., 2001. Exploring benefits of non-linear time compression. Proc. ACM Int. Conf. on Multimedia. ACM, Ottawa, Canada, pp. 382–391.

IEEE, 1969. IEEE recommended practice for speech quality measurement. IEEE Trans. Audio Electroacoust., 0018-9278 17 (3), 225–246. http://dx.doi.org/10.1109/TAU.1969.1162058.

Janse, E., 2004. Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech. Speech Commun. 42 (2), 155–173.

Janse, E., Nooteboom, S., Quené, H., 2003. Word-level intelligibility of time-compressed speech: prosodic and segmental factors. Speech Commun. 41 (2), 287–301.

Kawahara, H., Masuda-Katsuse, I., Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Commun. 27, 187–207, ISSN 0167-639.

Lebeter, J., Saunders, S., 2010. The effects of time compression on the comprehension of natural and synthetic speech. Working Papers Linguist. Circ. 20 (1), 63–81.

Maxemchuk, N., 1980. An experimental speech storage and editing facility. Bell Syst. Tech. J. 59 (8), 1383–1395. http://dx.doi.org/10.1002/j.1538-7305.1980.tb03370.x, ISSN: 0005-8580.

Moers, D., Wagner, P., Möbius, B., Müllers, F., Jauk, I., 2010a. Integrating a fast speech corpus in unit selection speech synthesis: experiments on perception, segmentation, and duration prediction. In: Proc. Speech prosody, Chicago, USA.

Moers, D., Jauk, I., Möbius, B., Wagner, P., 2010b. Synthesizing fast speech by implementing multi-phone units in unit selection speech synthesis. In: Proc. 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW-6).

Moos, A., Trouvain, J., 2007. Comprehension of ultra-fast speech – blind vs. 'normally hearing' persons. In: Proc. Int. Congress of Phonetic Sciences, vol. 1, pp. 677–680.

Neubarth, F., Pucher, M., Kranzler, C., 2008. Modeling Austrian dialect varieties for TTS. In: Proc. Interspeech, Brisbane, Australia, pp. 1877–1880

Nygaard, L., Pisoni, D., 1998. Talker-specific learning in speech perception. Percept. Psychophys. 60 (3), 355–376.

Port, R.F., 1981. Linguistic timing factors in combination. J. Acoust. Soc. Am. 69 (1), 262–274.

Pucher, M., Schabus, D., Yamagishi, J., 2010. Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and

sighted listeners. In: Proc. Interspeech, Makuhari, Japan, pp. 2186–2189.

Pucher, M., Toman, M., Schabus, D., Valentini-Botinhao, C., Yamagishi, J., Zillinger, B., Schmid, E., 2015. Influence of speaker familiarity on blind and visually impaired children's perception of synthetic voices in audio games. In: Proc. Interspeech, Dresden, Germany.

Rabiner, L., Schafer, R., 2010. Theory and Applications of Digital Speech Processing, first ed. Prentice Hall Press, Upper Saddle River, NJ, USA, ISBN 0136034284, 9780136034285.

Richmond, K., Clark, R., Fitt, S., 2010. On generating combilex pronunciations via morphological analysis. In: Proc. Interspeech, Makuhari, Japan, pp. 1974–1977.

Sanderman, A.A., Collier, R., 1997. Prosodic phrasing and comprehension. Language Speech 40 (4), 391–409.

Stent, A., Syrdal, A., Mishra, T., 2011. On the intelligibility of fast synthesized speech for individuals with early-onset blindness. Proc. ACM SIGACCESS Conference on Computers and Accessibility. ACM, Dundee, UK, pp. 211–218.

Syrdal, A.K., Bunnell, H.T., Hertz, S.R., Mishra, T., Spiegel, M.F., Bickley, C., Rekart, D., Makashay, M.J., 2012. Text-to-speech intelligibility across speech rates. In: Proc. Interspeech, Portland, USA.

Tachibana, M., Yamagishi, J., Masuko, T., Kobayashi, T., 2005. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. IEICE Trans. Inf. Syst. E88-D (11), 2484–2491.

Tokuda, K., Kobayashi, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In: Proc. ICASSP, Istanbul, Turkey, pp. 1315–1318.

Trouvain, J., 2007. On the comprehension of extremely fast synthetic speech. Saarland Working Papers Linguist. (SWPL) 1, 5–13.

Valentini-Botinhao, C., Toman, M., Pucher, M., Schabus, D., Yamagishi, J., 2014. Intelligibility analysis of fast synthesized speech. In: Proc. Interspeech, Singapore.

Verhelst, W., Roelands, M., 1993. An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. In: Proc. ICASSP, vol. 2, pp. 554–557.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J., 2009. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio, Speech Language Process. 17 (1), 66–83. http://dx.doi.org/10.1109/TASL.2008.2006647, ISSN: 1558-791.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 1998. Duration modeling for HMM-based speech synthesis. In: Proc. ICSLP, Sydney, Australia, pp. 29–32.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speaker interpolation for HMM-based speech synthesis system. Acoust. Sci. Technol. 21 (4), 199–206.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2007. A hidden semi-Markov model-based speech synthesis system. IEICE Trans. Inf. Syst. E90-D (5), 825–834, ISSN: 0916-8532.

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 52 (2010) 164–179
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis

Michael Pucher [a,*], Dietmar Schabus [a], Junichi Yamagishi [b], Friedrich Neubarth [c], Volker Strom [b]

[a] Telecommunications Research Center Vienna (ftw.), Donau-City-Str 1, 3rd floor, 1220 Vienna, Austria
[b] The Centre for Speech Technology Research, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, United Kingdom
[c] Austrian Research Institute for Artificial Intelligence (OFAI), Freyung 6/6, 1010 Vienna, Austria

Received 5 May 2009; received in revised form 22 September 2009; accepted 23 September 2009

Abstract

An HMM-based speech synthesis framework is applied to both standard Austrian German and a Viennese dialectal variety and several training strategies for multi-dialect modeling such as dialect clustering and dialect-adaptive training are investigated. For bridging the gap between processing on the level of HMMs and on the linguistic level, we add phonological transformations to the HMM interpolation and apply them to dialect interpolation. The crucial steps are to employ several formalized phonological rules between Austrian German and Viennese dialect as constraints for the HMM interpolation. We verify the effectiveness of this strategy in a number of perceptual evaluations. Since the HMM space used is not articulatory but acoustic space, there are some variations in evaluation results between the phonological rules. However, in general we obtained good evaluation results which show that listeners can perceive both continuous and categorical changes of dialect varieties by using phonological transformations employed as switching rules in the HMM interpolation.

Keywords: Speech synthesis; Hidden Markov model; Dialect; Sociolect; Austrian German

1. Introduction

Statistical parametric speech synthesis based on hidden Markov models (HMMs) (Yoshimura et al., 1999) has become established and well-studied, and has an ability to generate natural-sounding synthetic speech (Black et al., 2007; Zen et al., 2009). In recent years, the HMM-based speech synthesis systems have reached performance levels comparable to state-of-the-art unit selection systems (Fraser and King, 2007; Karaiskos et al., 2008). In this method, acoustic features such as the spectrum, excitation parameters, and segment duration are modeled and generated simultaneously within a unified HMM framework. A significant advantage of this model-based parametric approach is that speech synthesis is far more flexible compared to conventional unit selection methods, since many model adaptation and model interpolation methods can be used to control the model parameters and thus the characteristics of the generated speech (Yoshimura et al., 2000; Yamagishi et al., 2009a). In fact, these methods have already been applied to generating transitions between different speakers (Yoshimura et al., 2000), different types of emotional speech, and different speaking styles (Tachibana et al., 2005).

These techniques are also useful for achieving *varying* multi-dialect voices in text-to-speech (TTS) synthesis. They may be used for personalizing speech synthesis systems and have several potential benefits. For example, if the TTS system is used to provide an alternative voice output for

* Corresponding author.
E-mail address: pucher@ftw.at (M. Pucher).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.09.004

patients who have progressive dysarthria (Creer et al., 2009), some patients will desire a TTS system that has the same dialect as themselves.

However, it is not always feasible to prepare pronunciation dictionaries separately for every possible language variety in advance, since writing dictionaries is an extremely time-consuming and costly process. Often one variety is taken as a standard, and the linguistic resources such as pronunciation dictionaries are only available for this standard variety. Thus, to flexibly model as many varieties as possible, some acoustic and linguistic control based on this standard or typical dialect is required.

Although one might regard dialect control[1] as conceptually equivalent to the emotional control mentioned above, there is a significant difference in the requirements for the control of dialectal varieties. The speaker or emotional interpolation mentioned above implicitly assumes that the target models use the same pronunciation dictionary, and therefore phone strings, within the same language and linear interpolation is applied just to the relevant models, which results in acoustic transitions within the same phone or sub-word unit. For dialect control, we need to additionally consider linguistically-motivated transitions. In other words, we need to include not only the HMMs but also the pronunciation dictionary as targets of the interpolation process. That is, the HMMs to be interpolated may represent different phone sequences derived from different dictionaries. Moreover, these sequences may also consist of a different number of phones.

A major premise for dialect control is that dialects, as varieties of languages, form a "continuum" (Saussure, 1983): the varieties are related to one another in terms of being linguistically close, which makes it possible for us to hypothesize the existence of varieties on that continuum of fine-grained subtleties that lie between two different varieties already defined by linguistic resources. In addition to geographical transition of the dialect varieties, that is, regiolects, we may apply the same logic to other varieties of languages such as sociolects, which are categories of linguistic varieties defined by the social level of speakers.

The proposed dialect interpolation aims to produce synthetic speech in a phonetically intermediate variety from given models and dictionaries for adjacent typical varieties. For the phonetic control, we simply use linear interpolation of HMMs that represent the acoustic features similar to speaker or emotional interpolation. Since relations between articulatory and acoustic features are non-linear (Stevens, 1997), the phonetic control that can be achieved using acoustic features alone is noisy and might sometimes exhibit unexpected behavior. However it is worthwhile to investigate the basic performance of acoustic interpolation because proper acquisition of articulatory features requires specialized recording equipment such as electromagnetic articulography (EMA) (Schönle et al., 1987) and also because phonetic knowledge such as vowel height or backness and place or manner of articulation can be used in clustering the acoustic HMMs via manually-defined linguistic questions.

A closer inspection of potential phonetic transitions between language varieties reveals several exceptional cases. From phonetic studies of Viennese dialects (Moosmüller, 1987) we know that some gradual transitions are well motivated (e.g., spirantization of intervocalic lenis plosives), while some other transitions between phones are strong markers for that specific variety, and thereby categorical. In the latter case, either the standard form of a given phone is produced, or its dialectal counterpart, with no possible in-between variants. One example of such a transition is the phone [aː] in the standard Austrian German variety which is realized as [ɔː] in the Viennese dialect (mentioned in detail later in Table 5). For such a case, the use of interpolation (e.g., model interpolation between [aː] and [ɔː] phone HMMs) is not appropriate. For this reason, we introduce several knowledge-based switching rules that allow for overriding acoustic interpolation in such cases. Since it is known from psycholinguistics that continuous transitions between phones are often only perceived categorically (Liberman, 1970), the knowledge-based switching rules should improve the perception of dialects compared to acoustic interpolation alone. Hence, we include interpolations with and without switching rules in the subjective evaluation to measure the effect of the proposed dialect interpolation and switching rules.

In addition we investigate efficient clustering strategies for the dialect varieties in HMM-based speech synthesis. In general there are insufficient speech resources for non-standard dialect varieties. This situation might be even more severe for minor languages. Thus we compare several clustering algorithms for a practical case where the amount of speech data for dialects is limited, but there is sufficient speech data for the standard. We also include speech data from speakers that are able to speak standard and dialect.

This paper is organized as follows. Section 2 gives an overview of modeling strategies of HMM-based speech synthesizers for dialect varieties and an associated evaluation. In Section 3 we show how to generate speech that forms a continuous transition between one variety and another. The two varieties we are considering in this paper are standard Austrian German and Viennese dialect. Apart from continuous interpolation of HMMs, we also define specific switching rules. We then present the results of a series of listening tests. Section 4 summarizes our findings and discusses remaining issues.

[1] In this paper we use the notion of 'dialect' in a broad sense as referring to non-standard language varieties. In the case at hand, it would be more accurate to speak of Viennese sociolect, since language varieties in Vienna are discerned by social criteria and not (or no longer) identified by association to a certain geographical region. We use the term 'dialect control' as shorthand for 'control of dialectal or sociolectal language variety'.

2. Acoustic modeling of dialect varieties in HMM-based speech synthesis

2.1. Overview of HMM-based TTS system

All TTS systems described here are built using the framework from the "HTS-2007/2008" system (Yamagishi et al., 2009b; Yamagishi et al., 2008), which was a speaker-adaptive system entered for the Blizzard Challenges in 2007 and 2008 (Karaiskos et al., 2008). The HMM-based speech synthesis system, outlined in Fig. 1, comprises four main components: speech analysis, average voice training, speaker adaptation, and speech generation.

In the speech analysis part, three kinds of parameters for the Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram (STRAIGHT) (Kawahara et al., 1999) mel-cepstral vocoder (Tokuda et al., 1991, 1992) with mixed excitation (Yoshimura et al., 2001,) (i.e. a set including the mel-cepstrum, $\log F_0$, and band aperiodicity measures) are extracted as feature vectors for the HMMs. These features are described in (Zen et al., 2007a). In the average voice training part, context-dependent multi-stream left-to-right multi-space distribution (MSD) hidden semi-Markov models (HSMMs) (Zen et al., 2007b) are trained on multi-speaker databases in order to simultaneously model the acoustic features and duration. The phonetic and linguistic contexts we employ contain phonetic, segment-level, syllable-level, word-level, and utterance-level features as follows:

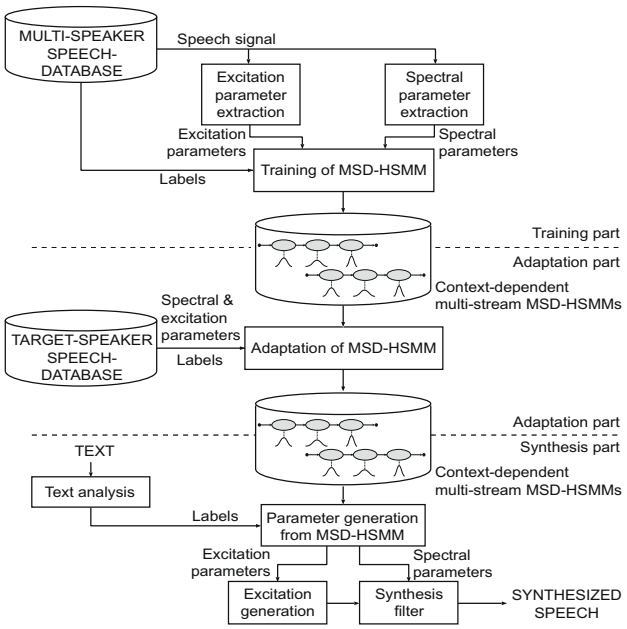


Fig. 1. Overview of the HTS-2007 speech synthesis system, which consists of four main components: speech analysis and feature extraction, average voice training on multi-speaker speech database, speaker adaptation to a target speaker, and speech generation from the adapted models (Yamagishi et al., 2009b; Yamagishi et al., 2008).

- preceding, current, and succeeding phones;
- acoustic and articulatory classes of preceding, current, and succeeding phones;
- the part of speech of the preceding, current, and succeeding words;
- the number of syllables in the preceding, current, and succeeding accentual phrases;
- the type of accent in the preceding, current, and succeeding accentual phrases;
- the position of the current syllable in the current accentual phrase;
- the number of accented syllables before and after the current syllable in the current phrase;
- the number of syllables in the preceding, current, and succeeding breath groups;
- the position of the current accentual phrase in the current breath group;
- the number of words and syllables in the sentence;
- the position of the breath group in the sentence;
- the specific language variety in the case of clustering of dialects (i.e. Viennese dialect or standard Austrian German).

Phonesets used for standard Austrian German and Viennese dialect are shown in Table 1. Austrian German and Viennese dialect have 58 and 75 phones, respectively. A set of model parameters (mean vectors and covariance matrices of Gaussian probability density functions (pdfs)) for the speaker-independent MSD–HSMMs is estimated using the feature-space speaker-adaptive training (SAT) algorithm (Anastasakos et al., 1996; Gales, 1998). In the speaker adaptation part, the speaker-independent MSD–HSMMs are transformed by using constrained structural maximum *a posteriori* linear regression (Yamagishi et al., 2009a).

In the speech generation part, acoustic feature parameters are generated from the adapted MSD–HSMMs using a parameter generation algorithm that considers both the global variance of a trajectory to be generated and trajectory likelihood (Toda and Tokuda, 2007). Finally, an excitation signal is generated using mixed excitation (pulse plus band-filtered noise components) and pitch-synchronous overlap and add (PSOLA) (Moulines and Charpentier, 1990). This signal is used to excite a mel-logarithmic spectrum approximation (MLSA) filter (Fukada et al., 1992) corresponding to the STRAIGHT mel-cepstral coefficients and thus to generate the speech waveform.

2.2. Speech database for Austrian German and Viennese dialect

For training and adaptation of Austrian German and Viennese dialect voices, a set of speech data comprising utterances from 6 speakers was used. Table 2 shows details of the speakers and number of utterances recorded for each. Here *AT* stands for standard Austrian German and *VD* for Viennese dialect. There are many differences between standard Austrian German and Viennese dialect

Table 1
Phone sets used in the experiments, represented with IPA symbols. The coding for 'Austrian German' is in accordance with the phonetic analysis presented in (Muhr, 2007), the coding for 'Viennese dialect' reflects our own analysis. Phones in brackets indicate that these are not really members of the native set.

Category	Austrian German	Viennese dialect
vowel	a aː (ɔː) eː e̞ (e̞ː) iː i oː o̞ uː u yː y øː ø̞	a aː ɔ̝ ɔ̝ː e eː ɛ ɛː i iː ɪ o oː u uː ʊ y yː øː œ œː
di-/monophthong/nasal	a͡e a͡o e͡ø̞ (æ̃ː) (œ̃ː) (ɔ̃ː)	æː ɒː œː e͡ɪ̯ o͡ɪ̯ u͡ɪ̯ ãː ɔ̝̃ ɔ̝̃ː ĩː æ̃ː õː
r-vocalized	e̞ɐ e̞ːɐ iːɐ iɐ oːɐ o̞ɐ uːɐ uɐ yːɐ yɐ øːɐ ø̞ɐ	ɔ̝ɐ ɔ̝ːɐ e̞ɐ e̞ːɐ iɐ iːɐ o̞ɐ o̞ːɐ ʊɐ ʊːɐ (yːɐ) øːɐ
schwa	ə ɐ	ə ɐ
plosive	b d g p t k	b d g β ð ɣ p t k
fricative	f v s ʂ ʃ ʒ ç x h	f v s sː ʃ ç x h
liquid/nasal/glide	ʀ l m n ŋ j	ʀ l l̩ m m̩ n n̩ ŋ ŋ̩ j
silence/pause/glottis	*'sil' 'pau'* ʔ	*'sil' 'pau'* ʔ

Table 2
Data sources used for training and adaptation of standard Austrian German (*AT*) and Viennese dialect (*VD*) HMM-based speech synthesis systems.

Speaker	Gender	Age	Profession	Number of utterances	
				AT utterances	*VD* utterances
HPO	M	≈60	Actor	219	513
SPO	M	≈40	Radio narrator	4440	95
FFE	M	≈40	Engineer	295	–
BJE	M	≈50	Actor	87	95
FWA	M	≈60	Language teacher	87	95
CMI	M	≈35	Singer	–	95

on many linguistic levels, which we have described previously in (Neubarth et al., 2008). All speakers are male speakers, of which five are native speakers of the Viennese dialect. As we can see from this table, the data sets widely vary in terms of the number of utterances, and whether they contain speech data from standard, dialect, or both. This is simply because these speech data sources were recorded for different purposes: some were recorded for unit selection synthesis test voices (*BJE, CMI, FWA*), one data set was recorded for a small unit selection voice (*FFE*), one was recorded for a large unit selection voice (*SPO*), and one was recorded for the adaptation and interpolation experiments described here (*HPO*). Ideally we should use a well-balanced larger speech database having equal amounts of data from standard Austrian German and Viennese dialect in terms of quantity and linguistic contexts mentioned in the previous section. However since such a well-balanced database is not available yet and there are always fewer resources for non-standard varieties, we explore the best modeling for both *AT* and *VD* from the available unbalanced database.

Our first goal was to evaluate which modeling approach works best to train Austrian German and Viennese voices for the speaker *HPO* since this speaker's data is phonetically balanced for both *AT* and *VD* and this enables the evaluation of several modeling strategies.

2.3. Modeling approaches

Table 3 defines the modeling approaches we used. SD and SI refer to speaker-dependent and speaker-independent modeling. Likewise we can consider dialect-dependent and dialect-independent modeling. For dialect-independent modeling, there are two possible approaches. The first is to add dialect information as a context for sub-word units and perform decision-tree-based clustering of dialects in the training of the HMMs. The second is to divide a set of speech data in both varieties uttered by one speaker into two subsets of speech data in different varieties uttered by two different pseudo speakers. In similar way to SAT estimation (Anastasakos et al., 1996; Gales, 1998) where acoustic differences between speakers are normalized for better average voice model training, we can normalize acoustical differences between varieties and can train a more canonical dialect-independent model. We call this training procedure "dialect-adaptive training". DD, DI, DC and DN refer to dialect-dependent, dialect-independent, dialect clustering and dialect-adaptive training, respectively. DM refers to "DC plus DN". In the table, the first column gives a short name for each modeling method, the second column gives the target dialect of the adaptation, the third column gives the number of utterances available, the fourth and fifth columns show the dependency on speaker or dialect, in which × means negative and √ means positive for each factor, and the sixth and seventh columns show training with or without clustering of dialects and dialect-adaptive training.

In the clustering of dialects, a new question that distinguishes Viennese from Austrian German data is added to a set of questions for the decision-tree-based clustering (Young et al., 1994) and minimum description length (MDL) based automatic node-splitting (Shinoda and

Table 3
Definitions of modeling approaches used. SD and SI refer speaker-dependent and speaker-independent modeling. DD, DI, DC, DN, and DM refer to dialect-dependent, dialect-independent, dialect clustering, dialect-adaptive training, and DC plus DN, respectively. × means negative and √ means positive for each factor.

Name	Target	# Utterance	Data Dependency		Dialect	
			Speaker	Dialect	Clustering	Normalization
SD–DD (*AT*)	*AT*	219	√	√	×	×
SD–DD (*VD*)	*VD*	513	√	√	×	×
SD–DI	*AT*/*VD*	732	√	×	×	×
SD–DC	*AT*/*VD*	732	√	×	√	×
SD–DN	*AT*/*VD*	732	√	×	×	√
SD–DM	*AT*/*VD*	732	√	×	√	√
SI–DD (*AT*)	*AT*	5128	×	√	×	×
SI–DD (*VD*)	*VD*	892	×	√	×	×
SI–DI	*AT*/*VD*	6020	×	×	×	×
SI–DN	*AT*/*VD*	6020	×	×	×	√

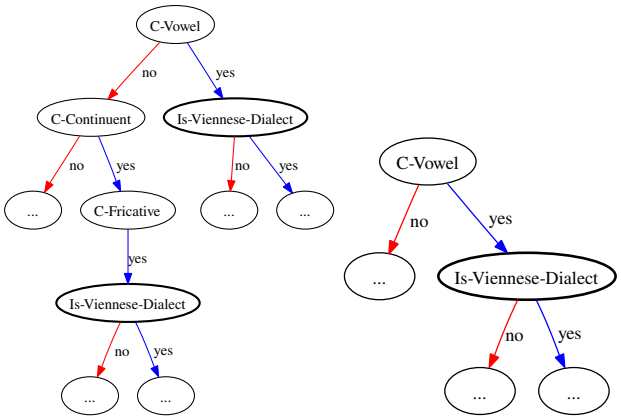


Fig. 2. Dialect clustering results. The left figure shows a part of a decision tree built for mel-cepstral coefficients and the right figure shows a part of a decision tree built for state duration. Both are for SD–DM systems.

Watanabe, 2000) is performed. Dialect is treated as a clustering context together with other phonetic and linguistic contexts and it is included in the single resulting acoustic model. Note that a decision tree was constructed independently for each combination of state index and acoustic parameter (mel-cepstrum, $\log F_0$, band aperiodicity) and duration. The same idea has been reported for multi-accented English average voice models (Yamagishi et al., 2008). In the clustering we observe that the question concerning the variety is used near the root of the decision trees. Fig. 2 shows part of the constructed decision tree for the mel-cepstral parameters of the third state and the corresponding duration parameter clustering tree. "C-Vowel" means "Is the center phoneme a vowel?", "C-Fricative" means "Is the center phoneme a fricative?", "Is-Viennese-Dialect" means "Is the current utterance in Viennese dialect?", and so on. From this example, we can see that separate Gaussian pdfs for vowel and fricative models for the Viennese dialect are produced from those for Austrian German. We can also see that separate Gaussian pdfs are generated for Viennese vowel duration.

We applied model adaptation with *AT* and *VD* data to all models except the first two. The adaptation represents dialect adaptation in the SD–DI, SD–DC, SD–DN, and SD–DM systems. It represents speaker adaptation in the SI–DD (*AT* or *VD*) systems. It represents simultaneous adaptation of speaker and dialect in the SI–DI and SI–DN systems. Therefore we have 16 voices in total (8 Austrian German and 8 Viennese voices), where 14 are adapted voices and 2 are speaker- and dialect-dependent voices.

2.4. Experimental conditions

Speech signals were sampled at a rate of 16 kHz and windowed by an F_0-adaptive Gaussian window with a 5 ms shift. The feature vectors per frame consisted of 138-dimension vectors: 39-dimension STRAIGHT mel-cepstral coefficients (plus the zeroth coefficient), $\log F_0$, 5 band-filtered aperiodicity measures, and their dynamic and acceleration coefficients. We used 5-state left-to-right context-dependent multi-stream MSD–HSMMs without skip transitions. Each state had a single Gaussian pdf with a diagonal covariance matrix in each stream for continuous features (mel-cepstra and band-limited aperiodicity) and MSDs consisting of scalar Gaussian pdfs and discrete distributions in each stream for log F_0 (Zen et al., 2007b) as emission probabilities, and also a Gaussian pdf as a duration probability. For speaker adaptation, the transformation matrices were triblock diagonal corresponding to the static, dynamic, and acceleration coefficients.

2.5. Evaluation

In order to choose the best voice for each variety that is used in the interpolation experiments in Section 3.2, a listening evaluation was conducted with 40 subjects. The listening evaluation consisted of two parts: in the first part listeners were asked to judge the overall quality of synthetic speech utterances generated from several models using the different training strategies from Table 3. The evaluation method used a 5-point scale, where 5 means "very good"

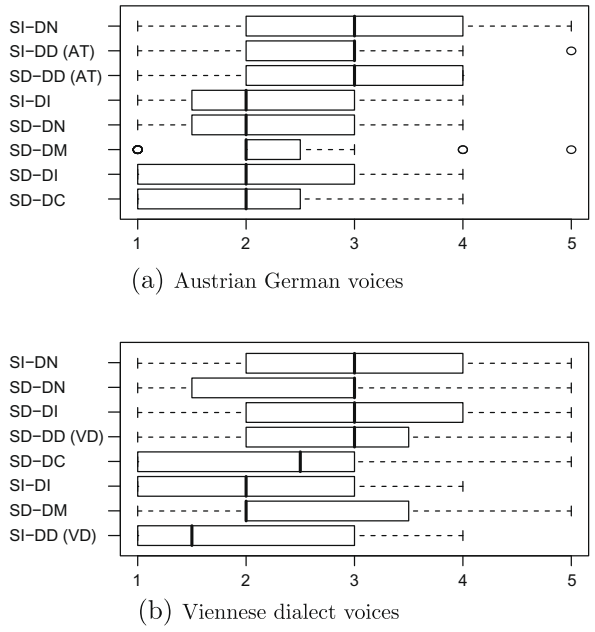


Fig. 3. Box-plots for 5-point scale evaluation for the overall quality for the *AT* (a) and *VD* (b) varieties. 5 means "very good" and 1 means "very bad". (a) Austrian German voices (b) Viennese dialect voices.

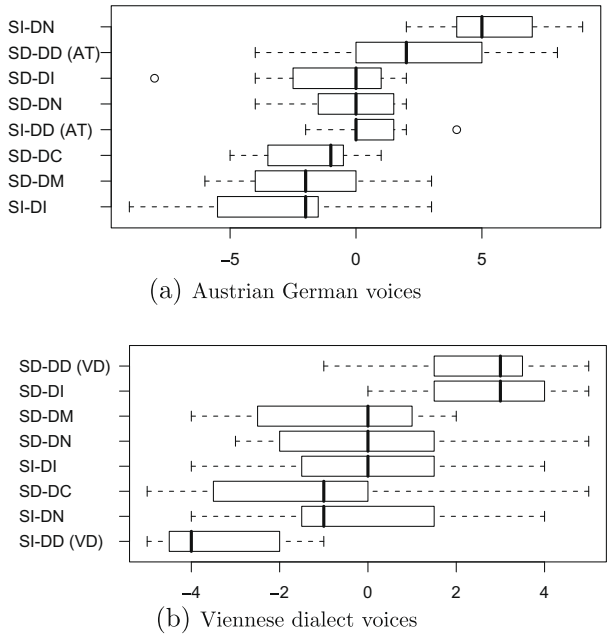


Fig. 4. Box-plots of pairwise comparison score for the *AT* (a) and *VD* (b) varieties. The data for one voice i comprise seven scores $s_j = w_{ij} - l_{ij}$, where $j \neq i$ and w_{ij} and l_{ij} are the numbers of comparisons won and lost, respectively, of voice i against voice j.

and 1 means "very bad". In the second part, after hearing a pair (in random order) of synthetic speech samples generated from the models, the listeners were asked which synthetic speech sample they preferred. The same synthetic speech utterances were used for both the evaluation tests. A Mann–Whitney–Wilcoxon test was used to identify significant differences.

Fig. 3 shows the results of the first part of the evaluation. For *AT*, there are three voices that are significantly better than other voices ($p < 0.05$), namely SI–DN, SI–DD (*AT*), and SD–DD (*AT*). For *VD*, the evaluation results for overall quality are less clear than those for the *AT* voices. Here we have only a clear loser with SI–DD, which is significantly worse than most other voices ($p < 0.05$) because of the low performance of the average voice model that was trained on a limited amount of *VD* speech data only.

These results are simply due to the amounts of data used for training the HMMs, rather than linguistic issues. In general, training of average voice models requires $O(10^3)$ utterances (Yamagishi et al., 2009a) but SI–DD (*VD*) has only about 900 utterances.

Fig. 4 shows the evaluation results for the pairwise comparisons of *AT* and *VD* voices. For the *AT* voices, the SI–DN voice is significantly better than all others except SD–DD (*AT*) ($p < 0.05$). However the speaker- and dialect-dependent SD–DD (*AT*) voice is significantly better than only two other voices; thus, the SI–DN voice may be considered the best. This is an interesting result: although we have enough *AT* speech data (particularly compared to *VD* speech data), the simultaneous use of both *AT* and *VD* speech data leads to better performance. This good performance of the adapted models is consistent with previous results (Yamagishi et al., 2009b). Furthermore we can see that the best training strategy is to divide utterances by a single speaker into standard (*AT*) and dialect (*VD*) utterances and treat them as two speakers in the SAT process, which is done in the SI–DN voice.

For *VD* there are two methods, namely SD–DD (*VD*) and SD–DI, that are significantly better than three other methods ($p < 0.05$). Since the speaker *HPO* has a relatively large amount of *VD* speech data but the amount of *VD* speech data from other speakers is very small, speaker-independent models do not perform well for *VD*.

From these results we chose SI–DN and SD–DD (*VD*) systems for the *AT* and *VD* voices, respectively. The *mixed variety* modeling approach is unfortunately not very successful, although we did observe some intuitively reasonable classes emerging from the clustering, such as a separate vowel cluster for the Viennese dialect. We believe that these problems are due to the limited amount of training data. We plan to repeat the experiments for the mixed dialect modeling when we have more balanced speech data available.

3. Dialect interpolation for HMM-based speech synthesis

In this section we add new phonological aspects to the model interpolation techniques for HMM-based speech synthesis, then apply this to dialect interpolation based on the concept of a dialect continuum. Specifically, we consider phonological rules which transform the standard

Table 4
Minor shifts between Austrian standard and Viennese dialect.

Phonological process	*AT* orthographic	Gloss	*AT* IPA	*VD* IPA
Tense vowels	**Bett**, offen	*bed, open*	bɐt, ofən	bet, ofɱ
Monophthongs	**Deutsch**	*German*	dɔ͡ɐtʃ	dætʃ
Spirantization	**Leber**, sorgen	*liver, worry*	leːbɐ, sɔɐgən	leːβɐ, sʊvɣŋ

Table 5
Phonologically-manifested differences of the Viennese dialect.

Phonological process	*AT* orthographic	Gloss	*AT* IPA	*VD* IPA
Input shift	**Schlag**, lieb	*cream, nice*	ʃlak, lip	ʃlɔːk, lɪɐp
l-vocalization-1	viele, **Keller**	*many, basement*	fiːlə, kɐlɐ	fyːlə, kœlɐ

variety to another variety. The rules between varieties determine which target phones are to be interpolated and the interpolation modes. In-between variants are thus generated using HMM interpolation under phonological constraints.

Differences between several typical English dialects are well-researched and well-formalized (e.g., Fitt and Isard, 1999). Certain differences between the standard variety of Austrian German and the Viennese dialect can also be formalized in phonological terms. Note that we are not concerned about differences on higher linguistic levels such as morphology – these have to be dealt with by generating different inputs and no direct comparison may be applied to them. We will first give an overview of the formalized phonological processes between the standard variety of Austrian German and the Viennese dialect.

3.1. Phonological processes between the standard variety of Austrian German and the Viennese dialect

The phonological differences between the language varieties under consideration can be classified according to formal criteria that also have a significant impact on the way one can interpolate between the models associated with different phones or phone strings (cf. Moosmüller, 1987; Neubarth et al., 2008):

1. *Minor shifts between Austrian standard and Viennese dialect* that are phonetically close and where these shifts are also observable in real life when people use different registers between the standard and some dialect variety (Table 4).
2. *Phonologically-manifested differences of the Viennese dialect* that are attributed to an 'input switch' between standard and dialect or differences that involve different phonological processes (Table 5).
3. *Differences affecting the segmental structure* by deleting or inserting phones from or into the phone string (Table 6).

The first set of differences involve vowels that only have a tense (closed, non-lowered) realization in the dialect variety, *monophthongization*, and the *spirantization* of inter-vocalic lenis plosives. The examples in Table 4 exemplify these processes. Crucially, the differences between the respective phones are gradual in a phonetic sense (Moosmüller, 1987). To model this group of processes and the transition between Austrian German and Viennese dialect, only an interpolation between phone models is necessary. Additionally, there are further common phonetically-motivated processes across word boundaries (hence post-lexical), which we did not consider in our experiments (assimilation of homorganic vowels, absorption of homorganic plosives, simplification of consonant clusters) (Moosmüller, 1987).

The second group of differences involve either different vowels (diachronically these phones have a different input base, so the notion *input shift* applies here), or different phonological processes apply to the input, while the segmental structure remains the same (Table 5). The term *l-vocalization(−1)* may be a little misleading here since the phone /l/ is not vocalized itself, but rather remains unchanged as an onset. However, it still spreads the feature (round) onto the preceding vocalic segment, and there are good reasons to view it as akin to the second version of *l-vocalization* (see below). Since the segmental structure is the same, it is unproblematic to apply a gradual interpolation between the relevant models – at least in a technical sense. For this kind of difference one normally does not find intermediate stages of a gradual shift in real life; rather, these differences are used to signal the use of a different (dialect) variety of a language. They are taken to be strong dialect markers. Depending on their presence or absence in an utterance or word it is perceived as dialect or not (Moosmüller, 1987).

The third group of differences shown in Table 6 poses a more difficult technical challenge, since the segmental structure changes. Most prominently these are instances of *l-vocalization* in non-onset position, where the phone /l/ forms a secondary rising diphthong with the preceding (round) vowel or is not realized at all, and various instances of *schwa-deletion*.

These groups of phonological processes may be applied in a combined fashion in order to achieve more complex phonological transitions between standard and dialects.

Table 6
Differences affecting the segmental structure.

Phonological process	*AT* orthographic	Gloss	*AT* IPA	*VD* IPA
l-vocalization-2	**Holz**, **Milch**	*wood, milk*	hɔ̯lts, milç	hɔ͡ɪts, myːç
Schwa-deletion	**Hände**, liege	*hands, lie*	hɛ̯ndə, liːgə	hent, lik
	Gewicht	*weight*	gəviçt	gviçt

Table 7
Applying processes selectively for the German word "Gefahr" (*'danger'*).

	AT IPA	*Process*		*VD* IPA
From *AT* to *VD*	[gəfaːr]	*schwa-deletion:*	[kfaːr]	[kfɔ̯ːɐ]
From *VD* to *AT*		*input shift /aː/*	[gəfɔ̯ːɐ]	

Table 7: *schwa-deletion* can be applied from the standard *AT* variety in order to indicate a slight approximation to the dialect without committing the speaker to strong dialect markers. Input shift for the vowel /a/ is always a strong dialect marker, but leaving the *schwa* pronounced indicates an approximation in the opposite direction, namely from dialect towards the standard. With this method it becomes possible also to model the direction of approximation between standard and dialect. In other words, it is possible to model a speaker of a certain variety who intends to speak another variety without fully committing him/herself to this variety.

3.2. Phonological constraints for HMM interpolation

For the first group mentioned in the previous subsection, we can straightforwardly apply HMM interpolation since they have the same number of phones in Austrian and Viennese. A good example is the distinction between di- and monophthongs in the Austrian Standard vs. Viennese dialect.

AT d ɶ͡e t ʃ
VD d æː t ʃ (1)

For simplification of HMM-level processing we assumed all phone HMMs have the same number of states and applied state-level interpolation in these experiments. Duration models for HSMM can also be interpolated. If the models had a different number of states, we would need to perform a state alignment between the two phone HMM sequences, based on some criterion (e.g., Kullback–Leibler divergence).

For the second group, which does not have in-between variants, we utilize simple switching rules which disable the HMM interpolation and switch the target phone for one variety to the other variety at some intermediate point (threshold). When such a threshold is given for the current phone, and the interpolation ratio for the utterance is below it, this phone is not interpolated, but rather the lower extreme point is used, as if the interpolation ratio were 0.0. If the interpolation ratio exceeds the threshold, the other extreme point (1.0) is used. Note that this is done phone by phone, so for neighboring phones it is possible that one is interpolated and the other is not.

This means that we can turn on or off the processes at a different point in the shifting continuum. Although we simply set this threshold to 0.5 in all our experiments, one could adjust this point for each phone individually.

For the third group (having words consisting of different numbers of phones in standard and dialect versions), we introduce a null phone [], which simply corresponds to a phone model with zero duration. Then, only the target phone's duration model is interpolated with the zero duration model.

AT g ə v i ç t
VD g [] v i ç t (2)

The above example (2) shows the alignment for the phonological process of *schwa-deletion* (Table 6) where the missing ə is aligned to the null duration model [].

Although these three groups and their combinations are not enough to automatically and completely reproduce the *VD* variety from the standard *AT* variety in TTS systems, we believe that they are sufficient to answer our scientific questions and to form a basis for our next large-scale experiments.

3.3. HMM linear interpolation and its underlying issues

From the above examples it should be clear that we cannot perform offline interpolation on the level of HMMs, since the same phone HMM may have several interpolation modes depending on what kinds of word the phone HMMs belong to and what kinds of phonological groups the word belongs to. Hence the interpolation of HMMs must be done on-line at synthesis time. We have therefore chosen *interpolation between observations* for the HMM interpolation, which was also used in (Tachibana et al., 2005) and is the simplest interpolation method described in (Yoshimura et al., 2000).

Fig. 5 shows the overall procedure flow for dialect interpolation. First we convert a given text into two context-dependent phoneme label sequences based on *AT* and *VD* pronunciation dictionaries. Then by consulting

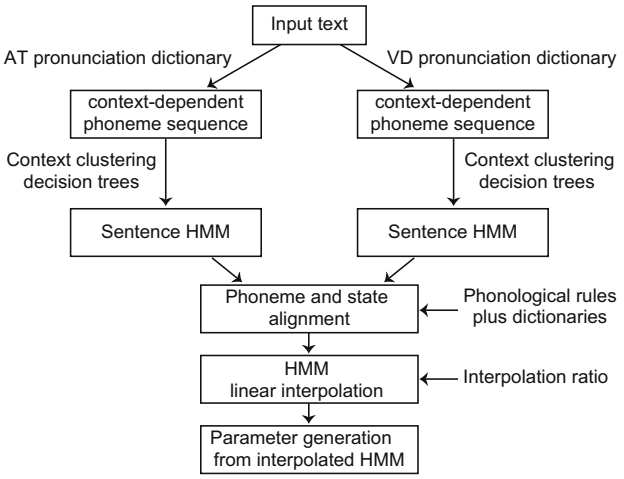


Fig. 5. Flow of dialect interpolation.

the context clustering decision trees built for each state of each feature in the HMMs for *AT* and *VD* voices separately, the context-dependent phoneme label sequences are converted into two sentence HMMs having different state sequences. Each state has several Gaussian pdfs for each of the acoustic features and a single Gaussian pdf for its duration. A Gaussian pdf for state i is characterized by a mean vector $\boldsymbol{\mu}_i$ and a covariance matrix $\boldsymbol{\Sigma}_i$. The dimension of the mean vector may vary depending on the acoustic features. Then, based on the pronunciation dictionaries and phonological rules adopted, the two state sequences are aligned and linear interpolation between the sequences is applied. Let $\boldsymbol{\mu}_i^{AT}$ and $\boldsymbol{\mu}_i^{VD}$ be mean vectors of Gaussian pdfs for *AT* and *VD* voices, respectively, at aligned state i. Likewise $\boldsymbol{\Sigma}_i^{AT}$ and $\boldsymbol{\Sigma}_i^{VD}$ are their covariance matrices. In the interpolation above (Yoshimura et al., 2000), the interpolated mean vector $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$ at state i are calculated as follows:

$$\hat{\boldsymbol{\mu}}_i = w\boldsymbol{\mu}_i^{AT} + (1-w)\boldsymbol{\mu}_i^{VD}, \tag{3}$$

$$\hat{\boldsymbol{\Sigma}}_i = w^2\boldsymbol{\Sigma}_i^{AT} + (1-w)^2\boldsymbol{\Sigma}_i^{VD}, \tag{4}$$

where w is an interpolation ratio between *AT* and *VD* voices. After all the Gaussian pdfs for all the acoustic features and their duration are interpolated in a similar way, an optimal acoustic trajectory is generated from the interpolated HMM sequence.

One obvious issue is that the HMMs represent acoustic features rather than articulatory features. Since the relationship between articulatory and acoustic features is non-linear (Stevens, 1997), it would be preferable to use articulator positions for the phonetic transition. In fact one of the authors and colleagues have already proposed "articulatory-controllable" HMM-based speech synthesis (Ling et al., 2008; Ling et al., 2009) based on this motivation. This would require the use of articulator positions; the current approach using only acoustic features is an approximation to this. Therefore it is expected that the current approach introduces some noise into the interpolation and may exhibit unexpected behavior from time to time. On the other hand, we emphasize that it is still worthwhile investigating the performance of such an acoustic interpolation, since proper acquisition of articulator positions requires specialized recording equipment. It is much easier to introduce phonetic knowledge such as vowel height or frontness and place or manner of articulation when clustering the acoustic HMMs via manually-defined linguistic questions.

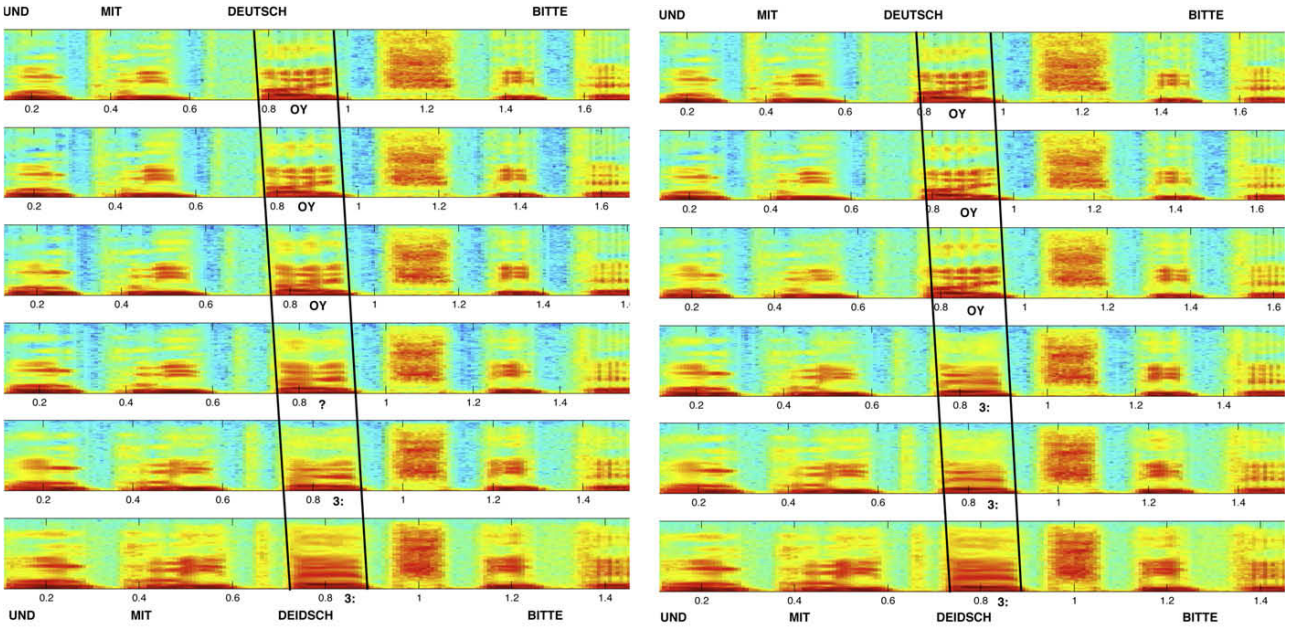
The success of the interpolation in the third group will also depend on whether the segment is the only vocalic portion of the syllable nucleus (as in the example *schwa-deletion* case above) or not. If it is the sole vocalic portion, intermediate stages may sound artificial because the vowel duration approaches zero and is thus too short to establish a phonetically-acceptable nucleus.

3.4. Interpolated examples

Fig. 6 shows spectrograms of synthetic speech interpolated between the *AT* variety (top) and the *VD* variety (bottom) in interpolation ratio increments of 0.2. In Fig. 6a only the HMM linear interpolation was used, whereas in Fig. 6b a combination of the HMM interpolation and switching rules was applied. These samples can be downloaded from *http://dialect-tts.ftw.at*. In Fig. 6a we can see the continuous transformation from /OY/ [ɶe] to /3:/ [æː]. Interestingly, while categorizing the sample utterances by experts, one intermediate stage was always classified as "undefined". This must be due to the non-linear relation between articulatory and acoustic features. In the other setting (Fig. 6b) a switching rule governs the application of either model for the relevant phone. The upper three spectrograms were generated with a model from Austrian Standard /OY/, the other lines with a model from Viennese dialect /3:/. The remaining parts of the utterance are interpolated linearly. This results in appropriate categorical transitions of phones. Fig. 7 shows the spectrogram of synthetic speech for the *schwa-deletion* case with and without switching rules. One can immediately see how the /@/ [ə] gradually disappears in Fig. 7a. All the intermediate stages except for the penultimate one are judged as sounding natural. In the one exception, the duration of the /@/ segment is too short to be either classified as completely missing or present. In Fig. 7b we can see the categorical transition of *schwa-deletion* with switching rules, which delete /@/. Gradual changes are possible for a set of phonological processes like *monophtongization* or *input shifts*, but they produce gaps in the acoustic perception with other processes like *schwa-deletion*. Additional samples are shown in Appendix A.

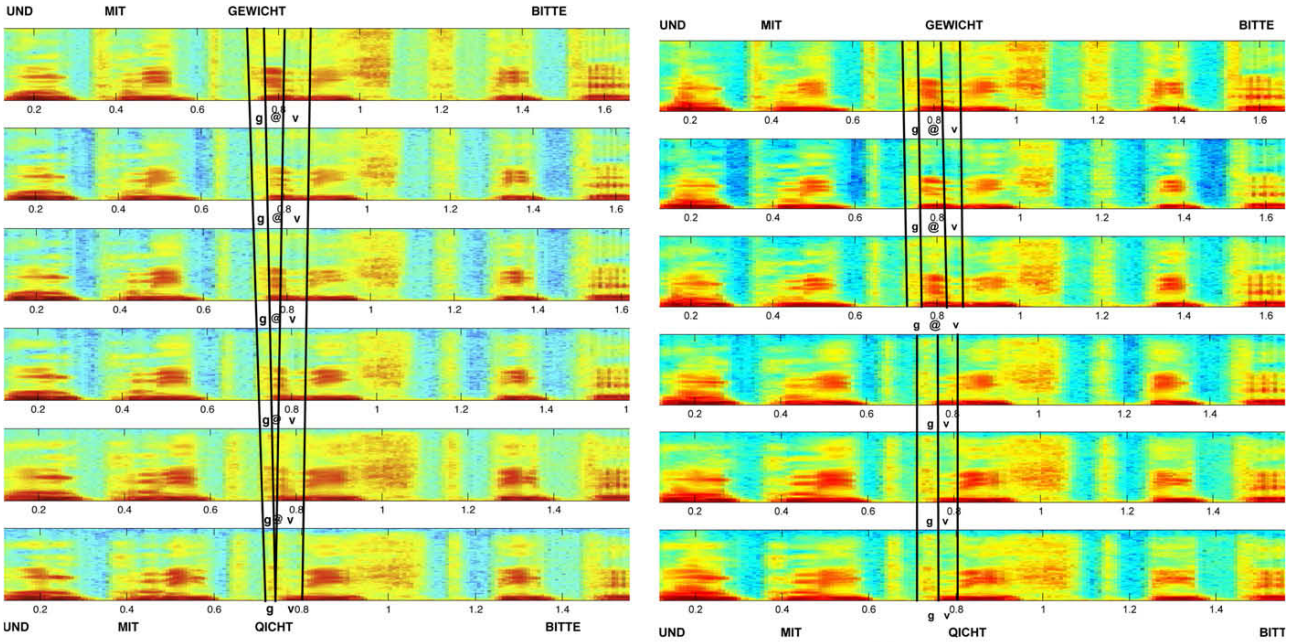
3.5. Evaluation

We designed a carrier sentence "Und mit ... bitte" (*And with ... please*) whose slot was filled with the words shown



(a) without switching rules (b) with switching rules

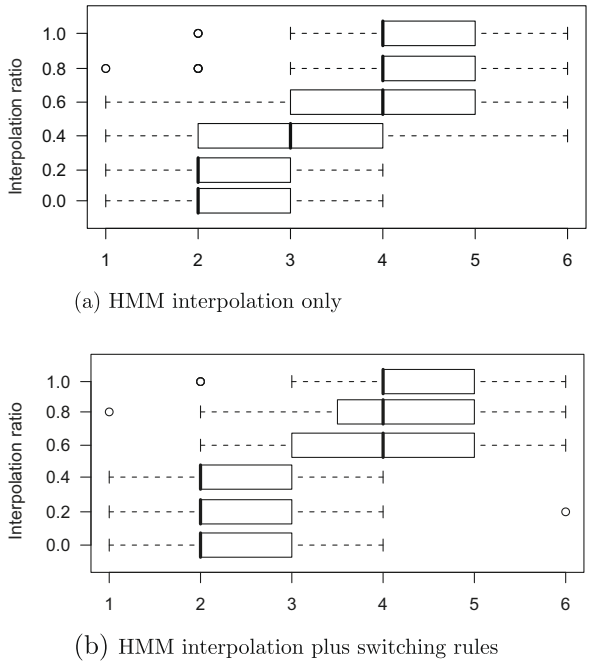
Fig. 6. An interpolation example between Austrian German "Und mit Deutsch bitte" (*And with German please*) and Viennese "Und mit Deidsch bitte". Interpolation ratio between them increments from 0.0 to 1.0 in steps of 0.2.



(a) without switching rules (b) with switching rules

Fig. 7. An interpolation example between Austrian German "Und mit Gewicht bitte" (*And with weight please*) and Viennese "Und mit Qicht bitte" having different segmental structures. Interpolation ratio between them increments from 0.0 to 1.0 in steps of 0.2.

in bold in Tables 4–6. Each word represents a different process, with the exception of *l-vocalization-1* and *schwa-deletion* which are used twice. The phonetic transcription of the carrier sentence is provided in example (5). This sentence has virtually no differences in different dialects.



(a) HMM interpolation only

(b) HMM interpolation plus switching rules

Fig. 8. Box-plot for all utterances. Interpolation (vertical axis) ranges from 0.0 to 1.0 with or without switching rule. On the horizontal axis, 1 means strongly *VD*, 6 means strongly *AT*.

AT/VD ʔ u n t m i t ... b i t ə (5)

For this evaluation we again used 40 listeners that had to answer two different questions after listening to synthesized interpolated prompts. In the first type of question, listeners were asked to give a rating as to what extent they would associate a given prompt with Viennese dialect or with standard Austrian German. For the rating, we used a scale from 1 (*'strongly Viennese'*) to 6 (*'strongly standard'*). Intermediate values were labelled *'Viennese'*, *'rather Viennese'* etc. In the second type of question, listeners were presented with two prompts and they were asked to judge how similar or different these were with respect to the differentiation between the dialect varieties. The first type of question is an *identification* task, the second type a *discrimination* task (Garman, 1990). The same Mann–Whitney–Wilcoxon test was used for finding significant differences.

Fig. 8 shows the overall results for the identification task. In the figure, a ratio of 0.0 corresponds to the *VD* non-interpolated speech samples and 1.0 corresponds to the *AT* non-interpolated speech samples. The interpolation ratio between them increments in step of 0.2; Fig. 8a shows results without switching rules and Fig. 8b shows results with switching rules applied to the phonological process. Overall we can see that a gradual change was perceived for the interpolations without switching rule and a categorical change was perceived with the interpolations that applied a switching rule relatively. The gradual change is underpinned by the significant differences between 0.2 and 0.4, between 0.4 and 0.6, and between 0.6 and 0.8. The categorical change due to the switching rules is

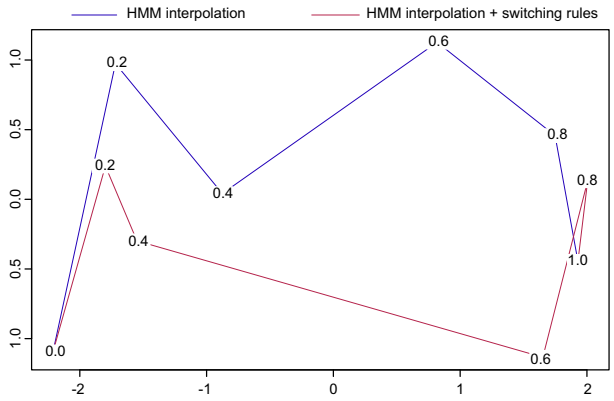


Fig. 9. Evaluation of similarity in terms of dialect. Multi-dimensional scaling is used for 2D visualization of evaluation results.

supported by the fact that there is a significant difference only between 0.4 and 0.6 ($p < 0.05$), and no significant differences between 0.2 and 0.4 or 0.6 and 0.8.

Fig. 9 shows the result of the discrimination task (pairwise comparison), visualized using multi-dimensional scaling (MDS) (Cox and Cox, 2001). From this figure, we can confirm several findings from the identification task. HMM interpolation generates continuous transitions: the first dimension found by MDS (horizontal axis) corresponds to this. Adding the switching rule causes this continuous transition to become categorical: 0.0, 0.2, and 0.4 are clustered at the left side and 0.6, 0.8 and 1.0 are clustered at the right side. There is a wide gap between 0.4 and 0.6 when the switching rules are applied. In fact, since the switch threshold was set to 0.5, the switching rule is applied between 0.4 and 0.6. The second dimension found by MDS (vertical axis) is related to the switching rules. Distances between switched and non-switched interpolations are represented by this dimension. Interpolated samples using a ratio of 0.6 with and without the switching rules are far apart: these samples were judged by the listeners to sound different. This is consistent with our earlier funding that experts always classified one intermediate stage as an undefined phoneme.

Fig. 10 shows the "Viennese-ness" ratings for three selected phonological processes, *monophthongization*, *input shift*, and *schwa-deletion* chosen from the three groups in Tables 4–6. We can clearly see the different behavior of these processes as dialect markers. The *monophthongization* process generates a relatively continuous transition between standard and dialect from both conditions. The *input shift* process generates a continuum between standard and dialect from the HMM interpolation, which does not match real phenomena, and generates a categorical shift with the switching rules. The *schwa-deletion* process creates a categorical shift at a certain point regardless of the use of switching rules. This means that a categorical change is perceived even if there is a continuous interpolation of the signal (Liberman, 1970).

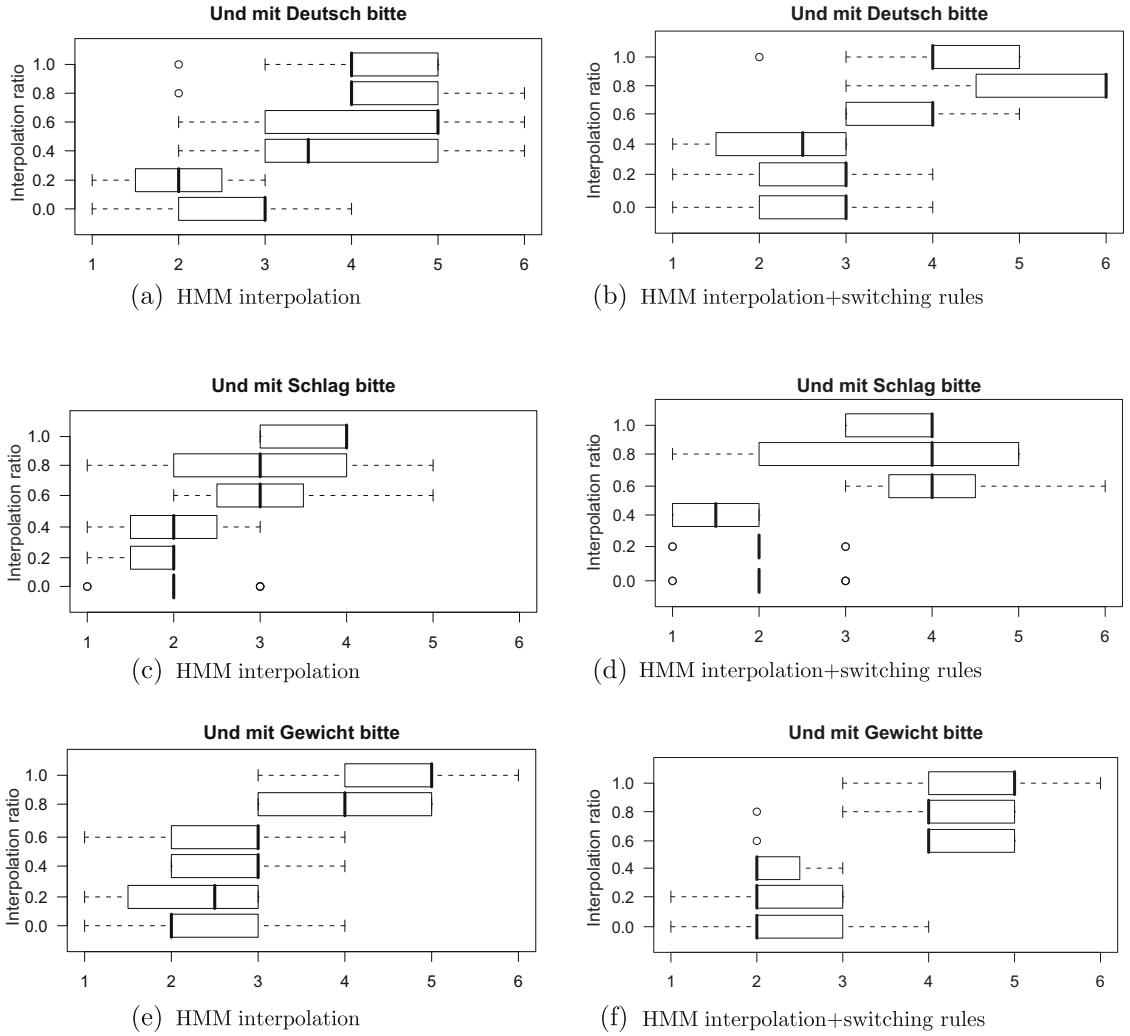


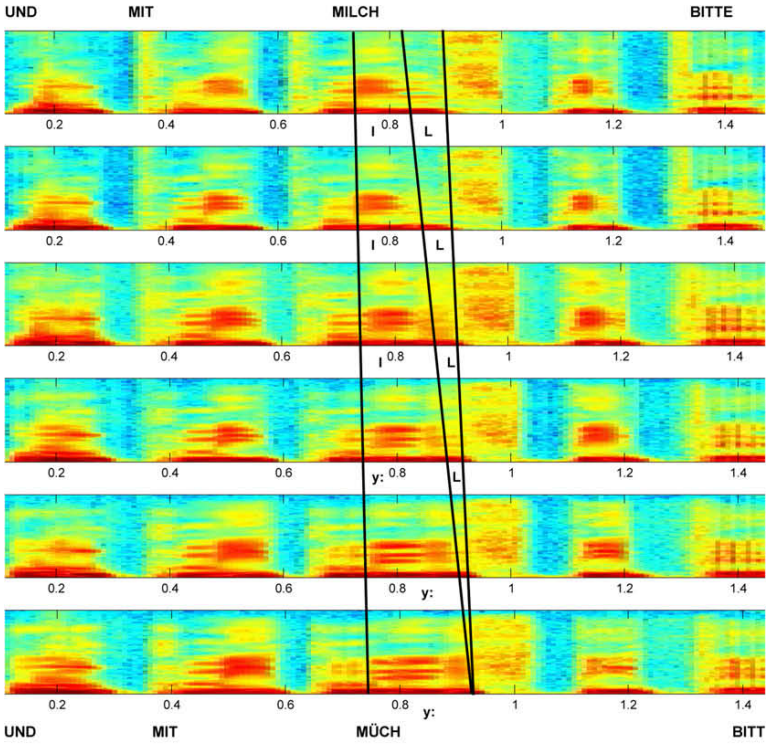
Fig. 10. Box-plots for three different utterances chosen from three categories. Interpolation (vertical axis) ranges from 0.0 to 1.0 with or without switching rule. On the horizontal axis, 1 means strongly *VD*, 6 means strongly *AT*.

4. Discussion and conclusion

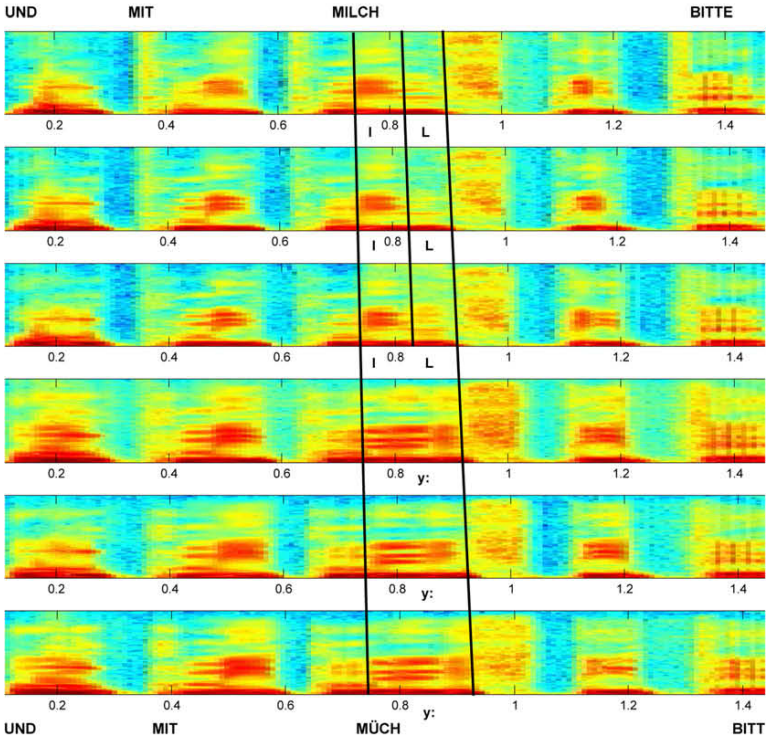
The HMM-based speech synthesis framework has been applied to Austrian German and Viennese dialect. We have investigated and evaluated several training strategies for multi-dialect modeling such as dialect clustering and dialect-adaptive training. Although the speech database was unbalanced in terms of the amount of Austrian German and Viennese dialect speech data, such a situation frequently occurs for non-standard varieties and so our results will apply to other dialects. For the *AT* variety, average voice models using dialect-adaptive training (where speech data uttered by a single speaker is divided into standard and dialect speaker data sets, and they are treated as different 'speakers' in the SAT process) achieve the best quality of synthetic speech. For the *VD* variety, speaker- and dialect-dependent modeling achieves the best quality. Although there was sufficient *AT* speech data, it did not help to improve the quality of the *VD* voice. We presume this is due to the linguistic differences between the *AT* and *VD* varieties.

In addition, we have bridged the gap between HMM-level processes and linguistic-level processes, by adding phonological processes to the HMM interpolation and applying it to dialect interpolation. We employed several formalized phonological rules between Austrian German and Viennese dialect as constraints for the HMM interpolation and verified their effectiveness in a number of perceptual evaluations. Since the HMM space used is not articulatory but simply acoustic, there are some variations in the effectiveness of each of the phonological rules. However, in general we obtained good evaluation results, which demonstrate that listeners can perceive both continuous and categorical changes of dialect variety in speech synthesised using phonological processes with switching rules in the HMM interpolation.

Our analysis results are obtained from relatively small-scale experiments designed to answer our scientific

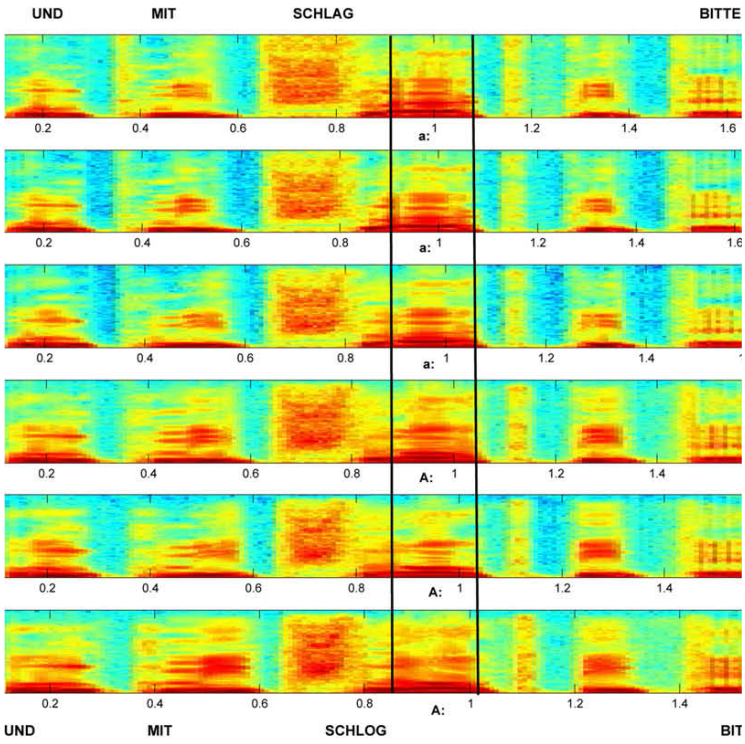


(a) without switching rules

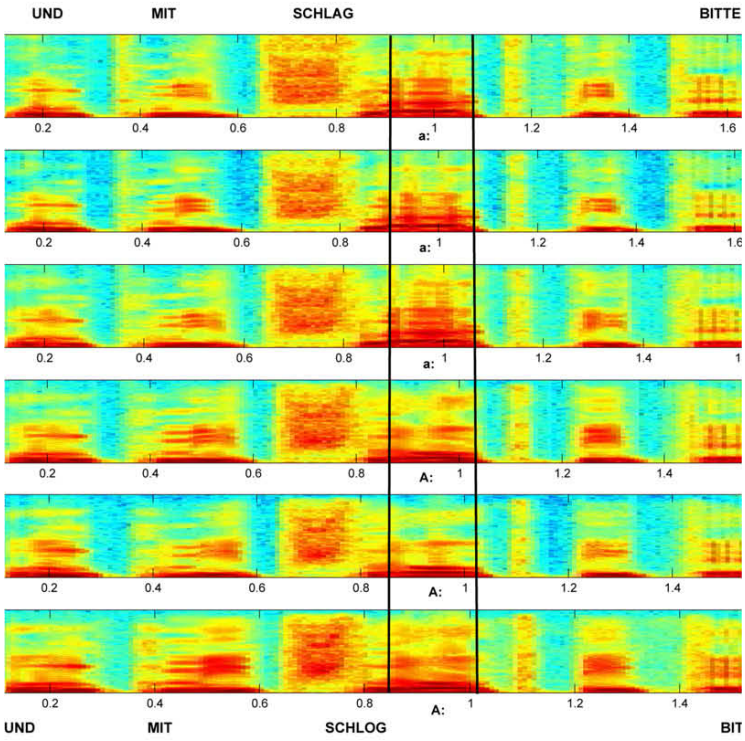


(b) with switching rules

Fig. 11. An interpolation example between Austrian German "Und mit Milch bitte" (*And with milk please*) and Viennese "Und mit Müch bitte" having different segmental structures. Interpolation ratio between them increments by 0.2.



(a) without switching rules



(b) with switching rules

Fig. 12. An interpolation example between Austrian German "Und mit Schlag bitte" (*And with cream please*) and Viennese "Und mit Schlog bitte". Interpolation ratio between them increments by 0.2.

questions and to form a basis for our future large scale experiments. For large scale experiments on automatic dialect interpolation, we need to identify and employ additional phonological rules for each dialect. More sophisticated models that use articulatory features may also bring improvements, especially for consonant transformation.

Our future work will also focus on an interpolation method that applies switching rules hierarchically to introduce the notion of direction into our modeling. Furthermore we wish to extend the interpolation strategy from the approach that uses a null phone to more sophisticated modeling approaches that use a distance metric on HMMs and dynamic programming to align sequences of models.

Acknowledgements

The project "Viennese Sociolect and Dialect Synthesis" is funded by the Vienna Science and Technology Fund (WWTF). The Telecommunications Research Center Vienna (FTW) is supported by the Austrian Government and the City of Vienna within the competence center program COMET. OFAI is supported by the Austrian Federal Ministry for Transport, Innovation, and Technology and by the Austrian Federal Ministry for Science and Research. Junichi Yamagishi is funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No. 213845 (the EMIME project). We thank Dr. Simon King and Mr. Oliver Watts of the University of Edinburgh for their valuable comments and proofreading. We also thank the reviewers for their valuable suggestions.

Appendix A. Additional interpolated examples

Fig. 11 shows the l-vocalization-2 process (Table 6) with and without switching rule. Without switching rule /L/ [l] gradually disappears and /I/ [ɪ] is gradually transformed into /y:/ [yː]. When a switching rule is applied /L/ is deleted.

Fig. 12 shows the *input shift* process (Table 5), which is very similar to *monophtongization* (Table 4) as shown in Fig. 6 where one vowel is transformed into another vowel. There is a continuous transformation when no switching rule is applied, whereas there is a categorical change when a switching rule is applied.

References

Anastasakos, T., McDonough, J., Schwartz, R., Makhoul, J., 1996. A compact model for speaker-adaptive training. In: Proc. ICSLP-96. pp. 1137–1140.

Black, A., Zen, H., Tokuda, K., 2007. Statistical parametric speech synthesis. In: Proc. ICASSP 2007. pp. 1229–1232.

Cox, T., Cox, M., 2001. Multidimensional Scaling. Chapman and Hall.

Creer, S., Green, P., Cunningham, S., 2009. Voice banking. Adv. Clin. Neurosci. Rehabil. 9 (2), 16–18.

Fitt, S., Isard, S., 1999. Synthesis of regional English using a keyword lexicon. In: Proc. Eurospeech 1999, Vol. 2. pp. 823–826.

Fraser, M., King, S., 2007. The Blizzard Challenge 2007. In: Proc. Blizzard 2007 (In: Proc. Sixth ISCA Workshop on Speech Synthesis), Bonn, Germany.

Fukada, T., Tokuda, K., Kobayashi, T., Imai, S., 1992. An adaptive algorithm for mel-cepstral analysis of speech. In: Proc. ICASSP-92. pp. 137–140.

Gales, M., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. 12 (2), 75–98.

Garman, M., 1990. Psycholinguistics. Cambridge University Press.

Karaiskos, V., King, S., Clark, R.A.J., Mayo, C., 2008. The Blizzard Challenge 2008. In: Proc. Blizzard Challenge Workshop, Brisbane, Australia.

Kawahara, H., Masuda-Katsuse, I., Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: possible role of a repetitive structure in sounds. Speech Commun. 27, 187–207.

Kawahara, H., Estill, J., Fujimura, O., 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In: 2nd MAVEBA.

Liberman, A.M., 1970. Some characteristics of perception in the speech mode. Percept. Disord. XLVIII (11).

Ling, Z.-H., Richmond, K., Yamagishi, J., Wang, R.-H., 2008. Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge. In: Proc. Interspeech, Brisbane, Australia. pp. 573–576.

Ling, Z.-H., Richmond, K., Yamagishi, J., Wang, R.-H., 2009. Integrating articulatory features into HMM-based parametric speech synthesis. IEEE Trans. Speech Audio Lang. Process. 17 (6), 1171–1185.

Moosmüller, S., 1987. Soziophonologische Variation im gegenwärtigen Wiener Deutsch. Franz Steiner Verlag, Stuttgart.

Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. 9 (5–6), 453–468.

Muhr, R., 2007. Österreichisches Aussprachewörterbuch Österreichische Aussprachedatenbank. Peter Lang Verlag, Frankfurt.

Neubarth, F., Pucher, M., Kranzler, C., 2008. Modeling Austrian dialect varieties for TTS. In: Proc. 9th Ann. Conf. Internat. Speech Communication Association (INTERSPEECH 2008), Brisbane, Australia. pp. 1877–1880.

Saussure, F.D., 1983. Course in General Linguistics. Duckworth, London, Original work published 1916.

Schönle, P.W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., Conrad, B., 1987. Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. Brain Lang. 31, 26–35.

Shinoda, K., Watanabe, T., 2000. MDL-based context-dependent subword modeling for speech recognition. J. Acoust. Soc. Jpn. (E) 21, 79–86.

Stevens, K., 1997. Articulatory-acoustic-auditory relationships. In: Hardcastle, W.J., Laver, J. (Eds.), The Handbook of Phonetic Sciences. Blackwell, Cambridge, pp. 462–506.

Tachibana, M., Yamagishi, J., Masuko, T., Kobayashi, T., 2005. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. IEICE Trans. Inf. Syst. E88-D (11), 2484–2491.

Toda, T., Tokuda, K., 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE Trans. Inf. Syst. E90-D (5), 816–824.

Tokuda, K., Kobayashi, T., Fukada, T., Saito, H., Imai, S., 1991. Spectral estimation of speech based on mel-cepstral representation. IEICE Trans. Fund. J74-A (8), 1240–1248, in Japanese.

Yamagishi, J., Zen, H., Wu, Y.-J., Toda, T., Tokuda, K., 2008. The HTS-2008 system: yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge. In: Proc. Blizzard Challenge 2008.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J., 2009a. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Speech Audio Lang. Process. 17 (1), 66–83.

Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S., Renals, S., 2009b. A robust speaker-adaptive HMM-based text-to-speech synthesis. IEEE Trans. Speech Audio Lang. Process. 17 (6), 1208–1230.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In: Proc. EUROSPEECH-99. pp. 2374–2350.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speaker interpolation for HMM-based speech synthesis system. Acoust. Sci. Technol. 21 (4), 199–206.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2001. Mixed excitation for HMM-based speech synthesis. In: Proc. EUROSPEECH 2001. pp. 2263–2266.

Young, S., Odell, J., Woodland, P., 1994. Tree-based state tying for high accuracy modelling. In: Proc. ARPA Human Language Technology Workshop, New Yersey, USA. pp. 307–312.

Zen, H., Toda, T., Nakamura, M., Tokuda, K., 2007a. Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. IEICE Trans. Inf. Syst. E90-D (1), 325–333.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2007b. A hidden semi-Markov model-based speech synthesis system. IEICE Trans. Inf. & Syst. E90-D (5), 825–834.

Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Commun. 51 (11), 1039–1064.

Influence of speaker familiarity on blind and visually impaired children's perception of synthetic voices in audio games

Michael Pucher[1], *Markus Toman*[1], *Dietmar Schabus*[1], *Cassia Valentini-Botinhao*[2]
Junichi Yamagishi[2,3], *Bettina Zillinger*[4], *Erich Schmid*[5]

[1] Telecommunications Research Center Vienna (FTW), Austria
[2] The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
[3] National Institute of Informatics, Japan
[4] University of Applied Sciences, Wiener Neustadt, Austria
[5] Federal Institute for the Blind, Vienna, Austria

{pucher,toman,schabus}@ftw.at, {cvbotinh,jyamagis}@inf.ed.ac.uk
bettina.zillinger@fhwn.ac.at, erich.schmid@bbi.at

Abstract

In this paper we evaluate how speaker familiarity influences the engagement times and performance of blind school children when playing audio games made with different synthetic voices. We developed synthetic voices of school children, their teachers and of speakers that were unfamiliar to them and used each of these voices to create variants of two audio games: a memory game and a labyrinth game. Results show that pupils had significantly longer engagement times and better performance when playing games that used synthetic voices built with their own voices. This result was observed even though the children reported not recognising the synthetic voice as their own after the experiment was over. These findings could be used to improve the design of audio games and lecture books for blind and visually impaired children.

Index Terms: speech perception, speech synthesis, audio games, blind individuals

1. Introduction

There is an ever increasing amount of applications that require customised speech synthesis that can reflect accent, speaking style and other features, particularly in the area of assistive technology [1, 2]. Current speech technology techniques make it possible to create synthetic voices that sound considerably similar to the original speaker using only a limited amount of training data [3]. This naturally leads to research questions regarding how a listener's perception of a synthetic voice depends on the listener's acquaintance with the speaker used to train the voice. Moreover how does one perceive a synthetic voice trained on one's own speech. These questions are particularly of interest when considering the design of audio lecture material for blind children and how learning may be improved by using familiar voices. One idea we are looking to exploit is the impact of using the child's own voice or that of their teacher.

To the best of our knowledge there are no existing studies on the perception of one's own synthetic voice. Studies on the perception of one's own natural voice exist but are quite sparse and do not report on preference or intelligibility results [4–6]. There is however an extensive literature on the perception of familiar voices [7–14]. Most studies create familiarity by exposing their listeners to a certain voice, either in one or a few sessions across a certain time range [10–12]. Such studies found that for both young adults [10,11] and older adults [12] prior exposure to a talker's voice facilitates understanding. In fact it's argued that this facilitation occurs because familiarity eases the effort for speaker normalization, i.e. the mapping of an acoustic realization produced by a certain speaker to a phonetic representation [15]. Relatively few studies evaluated the impact of long-term familiarity, i.e., a voice you have been exposed to for weeks, months or years [13, 14]. Newman and Evers [13] report an experiment of pupils shadowing a teacher's voice in the presence of a competing talker. Results show that pupils that were made aware that the target voice was their teacher's outperformed pupils that were unaware of this or that were unfamiliar with that particular teacher. Souza and colleagues [14] measured the long-term familiarity impact on speech perception by selecting spouses or pairs of friends and measuring how well they understand each other in noise. They found that speech perception was better when the talker was familiar regardless of whether the listeners were consciously aware of it or not.

There are also studies on the effect of familiarity of synthetic voices using a variety of synthesisers [16]. It has been shown that increased exposure to synthetic speech improves its process in terms of reaction time [16]. There are far fewer studies on the perception of synthetic speech which is similar to a particular person's voice or that has been synthesized with a particular voice [17, 18]. [17] showed that synthetic voices that are acoustically similar to one's own voice are generally not preferred over non-similar voices. A preference was however found for voices that showed the same personality as defined by duration, frequency, frequency range, and loudness of the voice. Another study [18] showed that it is more difficult for listeners to judge whether two sentences are spoken by the same person if one of the sentences is produced by a speech synthesizer and the other is natural speech as opposed to both being synthetic speech.

It has been shown that blind individuals obtain higher intelligibility scores when compared to sighted individuals [19] and that this benefit is also observed for the intelligibility of synthetic speech [20, 21] possibly due to the familiarity effect [22] as blind individuals are exposed to the material more through the use of screen readers and audio books.

In the context of a research project together with a school for blind children we evaluated the use of different synthetic voices in audio games. Assuming that synthetic voices still

Figure 1: *Studio recordings of blind school children.*

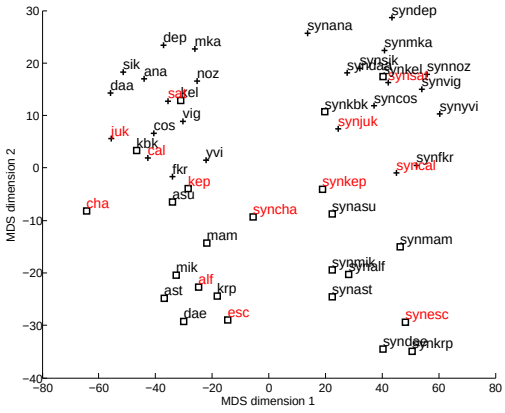


Figure 2: *Comparison between synthetic ("syn" affixed) and natural voices. School children are marked in black, teachers in red. Female speakers with crosses, male speakers with squares.*

benefit from the familiarity effect and that one's own synthetic voice is in a certain way a familiar voice, we evaluate the engagement time and game performance of a group of blind children playing audio games incorporating their own synthetic voice, their teacher's synthetic voice and an unknown synthetic voice. Using a HMM-based speech synthesis system for German we built voices of 18 school children and 7 teachers of the same school and an additional speaker who was not known to the children.

This paper is organised as follows: in Section 2, we describe the natural speech database used to train the voices and how they were created. In Section 3, we explain the design of the games, how to play them and measure their performance followed by Section 4 where we present experimental conditions and results. Finally, in Sections 5 and 6 we discuss our findings and conclude.

2. Speech databases and voices

To develop synthetic voices for the 18 children and 7 teachers of the school we recorded 200 phonetically balanced sentences for each speaker. The recordings were performed in an anechoic room with a professional microphone and recording equipment. Figure 2 shows the recording setup. For the blind children and teachers the sentences were played to the listeners via loudspeakers at a normal rate. We also recorded speech at fast and slow speaking rates from the same speakers. However these were not used in the current experiments. For the unfamiliar speaker's voice we used the same 200 sentences to develop a synthetic voice of the same quality as the children and teacher's.

When developing a synthetic voice for a speaker, we train a separate model for F0, spectrum, and duration for that speaker. These parameters are predicted for each speech unit by taking a large context into account. This leads to a more similar voice than only modifying certain speech parameters like overall duration, F0, frequency range, and loudness.

Figure 2 shows the comparison between all voices (natural and synthetic).To visualize the voices in a two-dimensional space we performed Dynamic Time Warping (DTW) between the same prompts from different speakers. For each of the 50 speakers (natural and synthetic) we had 29 different test prompts that were not used for voice training. Each prompt from a certain speaker was compared to the same prompt from all other speakers and the score was added to the respective speaker-speaker score. To obtain a similarity matrix for Multidimensional Scaling (MDS) we symmetrized the DTW scores. DTW uses the L_2 norm as distance metric.

Figure 2 shows the reduced two-dimensional space using only the two most significant dimensions. Along the horizontal axis we can see a speech type separation into a natural (left) and a synthetic (right) class. The vertical axis shows separation in terms of speaker. On this axis we can see that a certain speaker is closest to his/her respective synthetic voice. Furthermore, the y-axis shows a separation between female (crosses) and male (squares) speakers. Finally, there is no visible clustering according to age in this comparison as teachers (red font) are distributed across the space.

3. Audio games

To keep the children engaged with the experiment for a whole day, we integrated our experiments in audio-only games using speech synthesis. We developed two audio games to measure the impact of the chosen voice on game performance and engagement time.

3.1. Labyrinth game

The labyrinth game was used to measure engagement time. When starting up, instructions were presented to the player by the game voice. After the instructions, the player could choose between different labyrinth sizes: small with 7 rooms, medium with 15 rooms, large with 50 rooms and huge with 100 rooms. Keyboard cursor keys were used to navigate through the labyrinth, space bar allowed to replay the last spoken instruction, F1 presented help information to the user and F2 and F3 could be used to change the speaking rate of the game voice. The goal for the player was to find the exit of the labyrinth with as few steps as possible by remembering already visited rooms and labyrinth structure. The labyrinths were internally represented by randomly generated graphs with all nodes having a degree smaller than 4, a defined start and end point and a defined number of additionally attached dead ends. While the trees were randomly generated, the random seed used was the same for all players to ensure the experience would be the same for each player for each labyrinth size. Each node was randomly

assigned a room name (e.g., “kitchen”, “barn”) which was read to the player as well as the possible movement options (e.g., “You are now in the cockpit. Press left to go to the barn, press right to go to the kitchen.”) along the edges. Apart from the synthesized speech, non-disruptive ambient sounds were used as well as foot step sounds when moving through the labyrinth.

3.2. Memory game

The memory game was used to measure the performance of the player. As with the labyrinth game, when starting up, instructions were presented to the player by the game voice. Each round had a specific topic, e.g., musical instruments or animals. The game then constructed a non-visual, board with 8 (large: 16) fields and 4 (large: 8) items with each item associated with two fields (e.g. the item “elephant” was associated with the field belonging to keys a and j). A single key on a keyboard with German layout was associated with each field: a, s, d, f, j, k, l, ö for the normal field. For the large field, additional keys were added: q, w, e, r, u, i, o, p. Each turn consisted of the player being asked to press a key for the first field. Upon key press, the synthetic voice pronounced the item associated with the field. The player was then asked to pick a second field by pressing a key. Again upon selection, the synthetic voice pronounced the item associated with the field. If both fields were associated with the same item, the fields were removed from the current round. This was repeated until all duplicate items were found and all fields removed. Apart from the synthesized speech giving feedback on the player choices, sound effects were used for success or failure or pressing an invalid or already selected/removed key. At the end of each round, the player was told how many guesses he/she had needed to clear the board.

4. Experiments

For the experiments, 27 children played the two audio-only games. The children were grouped into 3 groups, where one group listened to their own synthetic voices in the games, one group listened to the teacher’s voices, and one group heard an unknown synthetic voice. For the children listening to the teacher’s voice we made sure that they knew the teacher very well from the classroom. Availability of a voice model, age (see Figure 5), gender and degree of visual impairment were the factors used to balance the groups. Note that it is, however, impossible to perfectly balance all the factors because of the limited number of blind children and their additional disabilities and hence we have used the three most balanced groups that we could define (see Figure 3).

The experiment was conducted in two computer rooms in school with the groups evenly split between the rooms. The games were deployed to the computers so that each child got a personalized version. They assumed that all of them were playing the same version of the game.

Figure 3 shows the descriptions of the users that participated in the experiment. We had 27 school children that participated in the evaluation. The users of speech synthesis and Braille displays were identical to the blind participants. Speakers were familiar with speech synthesis but not with HMM-based speech synthesis. We had slightly more female and blind participants in the first group.

Figure 4 shows the number of years blind users have been using speech synthesis technology and Braille displays. We can see that the blind children start to use Braille displays much earlier than speech synthesis.

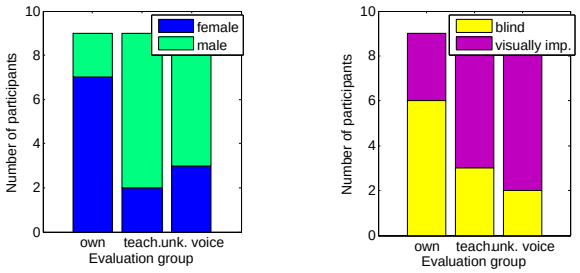


Figure 3: *Participants characteristics within groups.*

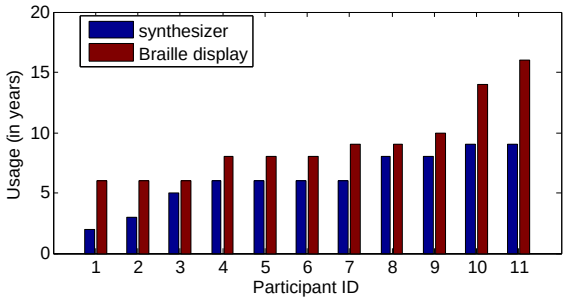


Figure 4: *Speech synthesis usage (blue bars) and Braille display usage (red bars) in years for the 11 blind participants.*

4.1. Labyrinth game

To measure engagement in the labyrinth game we used the time played overall and the number of games that were played. Children could choose how many games they wanted to play, and they could also choose the labyrinth size. The labyrinth game has a goal, namely finding the exit of the labyrinth, but it can also be played in an exploratory style where the players explore the rooms of the labyrinth.

Figure 6 (left) shows that participants hearing their own synthetic voice played significantly longer than users listening to an unknown synthetic voice ($p < 0.05$) according to a Wilcoxon rank sum test for equal medians. Differences between the teacher’s voice and unknown as well as own voices were not significant. The same trends are seen for groups with blind-only participants as shown in Figure 6 (right), but they are not significant. We did not find any significant gender differences for the labyrinth game.

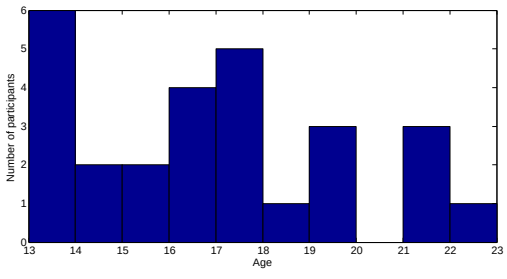


Figure 5: *Participants age distribution.*

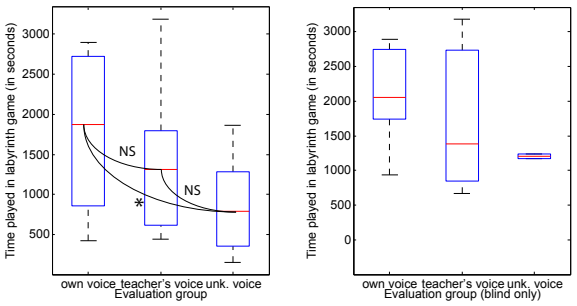


Figure 6: *Time played per group in the labyrinth game for all participants (left) and blind-only participants (right).*

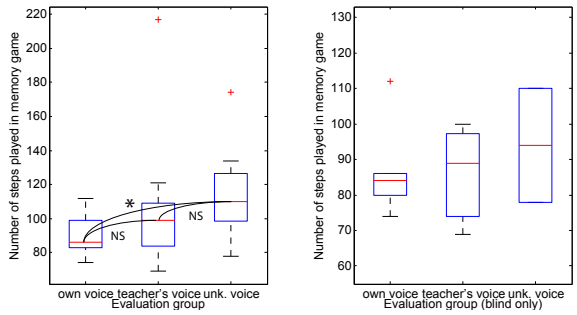


Figure 7: *Number of steps per group in the memory game for all participants (left) and blind-only participants (right).*

4.2. Memory game

In the experiments with the memory games the children had to play 8 mandatory rounds. As the conditions were the same for all children in this case, the first 6 rounds were on a normal game board, the next 2 on a large board. All children had the same topics for each round and the same assignments of items to fields. After playing the 8 rounds they could continue playing as long as they liked and freely choose the board size. To analyse the performance we only considered the 8 mandatory rounds. We used the number of steps needed to solve all 8 rounds as performance variable.

Figure 7 (left) shows that the children needed significantly less steps ($p < 0.05$) for finishing the memory game when using their own synthetic voice compared to an unknown synthetic voice. Differences between the teacher's voice and unknown as well as own voices were not significant. Again we can see the same trends also for groups with blind-only participants, but they are not significant. No significant gender differences were found for the memory game.

4.3. Blind vs. visually impaired users

As Figure 8 shows, blind participants played significantly longer ($p < 0.05$) than visually impaired participants. This is true for the labyrinth as well as for the memory game. The stronger engagement of blind users in playing is also true for other performance variables. We think that blind users are more sensitive to the auditive modality and can thereby gain more pleasure in playing audio-only games.

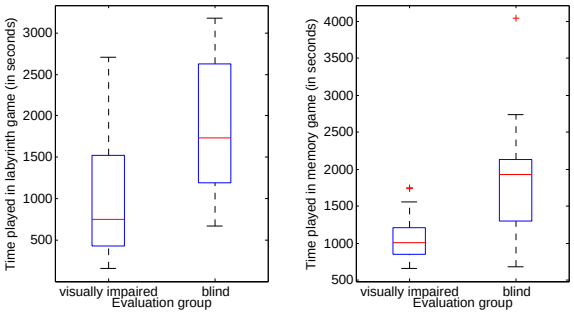


Figure 8: *Time spent playing the labyrinth game (left) and memory game (right) for blind vs. visually impaired players.*

5. Discussion

Our results show that the use of one's own voice increases the engagement time in audio games, which indicates a certain preference. To align our results with the results in [17] one's own voice can also be considered as the extreme case of a voice from a speaker with the same personality as oneself. Results for listeners of teacher's voices, although not significant, show a trend that reflects the special role of familiarity when a voice of a speaker to which the listener has a special social relation (teacher) is concerned.

The children in our study prefer known voices although they did not recognise the speakers (neither themselves nor the teachers). This indicates a certain type of cognitive processing where speech recognition and speaker recognition are independent but features of familiar speakers can be used in the recognition process. This ease of recognition of familiar speakers could be one explanation for the longer engagement times.

6. Conclusion

In this paper, we have shown that listening to one's own synthetic voice increases engagement and performance of blind school children in audio games significantly. For the evaluation we developed an audio-only labyrinth game to measure engagement time and a memory game to measure performance. Familiar voices like teacher's voices show a trend of increased engagement and performance, but more experiments are needed for verifying this hypothesis.

We also showed that blind listeners engage longer with the audio games than visually impaired listeners. We hypothesize that blind listeners are more accustomed to listening to synthetic speech and it is easier for them to process synthetic speech.

For blind users that are using speech synthesis on a regular basis there is a need to make their synthesizer experience more engaging and pleasurable, which can be accomplished by using their own or familiar voice in the synthesizer.

7. Acknowledgement

This work was supported by the BMWF - Sparkling Science project *Sprachsynthese von Auditiven Lehrbüchern für Blinde SchülerInnen* (SALB) and by the Austrian Science Fund (FWF): P23821-N23. The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH is funded within the program COMET - Competence Centers for Excellent Technologies by BMVIT, BMWFJ, and the City of Vienna. The COMET program is managed by the FFG.

8. References

[1] M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, and V. Strom, "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis," *Speech Communication*, vol. 52, no. 2, pp. 164–179, 2010.

[2] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, 2012.

[3] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[4] C. Fernyhough and J. Russell, "Distinguishing one's own voice from those of others: A function for private speech?" *International Journal of Behavioral Development*, vol. 20, no. 4, pp. 651–665, 1997.

[5] R. Appel and J. G. Beerends, "On the quality of hearing one's own voice," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 237–248, 2002.

[6] C. Rosa, M. Lassonde, C. Pinard, J. P. Keenan, and P. Belin, "Investigations of hemispheric specialization of self-voice recognition," *Brain and cognition*, vol. 68, no. 2, pp. 204–214, 2008.

[7] D. Van Lancker, J. Kreiman, and K. Emmorey, "Familiar voice recognition: Patterns and parameters. part i: Recognition of backward voices." *Journal of phonetics*, vol. 13, pp. 19–38, 1985.

[8] D. V. Lancker and J. Kreiman, "Voice discrimination and recognition are separate abilities," *Neuropsychologia*, vol. 25, no. 5, pp. 829–834, 1987.

[9] T. Bőhm and S. Shattuck-Hufnagel, "Utterance-final glottalization as a cue for familiar speaker recognition," in *Proc. Interspeech, Antwerp*, 2007, pp. 2657–2660.

[10] L. C. Nygaard, M. S. Sommers, and D. B. Pisoni, "Speech perception as a talker-contingent process," *Psychological Science*, vol. 5, no. 1, pp. 42–46, 1994.

[11] L. C. Nygaard and D. B. Pisoni, "Talker-specific learning in speech perception," *Perception & psychophysics*, vol. 60, no. 3, pp. 355–376, 1998.

[12] C. A. Yonan and M. S. Sommers, "The effects of talker familiarity on spoken word identification in younger and older listeners." *Psychology and aging*, vol. 15, no. 1, p. 88, 2000.

[13] R. S. Newman and S. Evers, "The effect of talker familiarity on stream segregation," *Journal of Phonetics*, vol. 35, no. 1, pp. 85 – 103, 2007.

[14] P. Souza, N. Gehani, R. Wright, and D. McCloy, "The advantage of knowing the talker," *Journal of the American Academy of Audiology*, vol. 24, no. 8, p. 689, 2013.

[15] D. Pisoni and R. Remez, *The Handbook of Speech Perception*. John Wiley & Sons, 2008.

[16] M. Reynolds, C. Isaacs-Duvall, B. Sheward, and M. Rotter, "Examination of the effects of listening practice on synthesized speech comprehension," *Augmentative and Alternative Communication*, vol. 16, no. 4, pp. 250–259, 2000.

[17] C. Nass and K. M. Lee, "Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction." *Journal of Experimental Psychology: Applied*, vol. 7, no. 3, p. 171, 2001.

[18] M. Wester and R. Karhila, "Speaker similarity evaluation of foreign-accented speech synthesis using HMM-based speaker adaptation," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 5372–5375.

[19] K. Hugdahl, M. Ek, F. Takio, T. Rintee, J. Tuomainen, C. Haarala, and H. Hmlinen, "Blind individuals show enhanced perceptual and attentional sensitivity for identification of speech sounds," *Cognitive Brain Research*, vol. 19, no. 1, pp. 28 – 32, 2004.

[20] K. Papadopoulos, V. S. Argyropoulos, and G. Kouroupetroglou, "Discrimination and comprehension of synthetic speech by students with visual impairments: The case of similar acoustic patterns," *Journal of Visual Impairment & Blindness*, vol. 102, no. 7, pp. 420–429, 2008.

[21] M. Pucher, D. Schabus, and J. Yamagishi, "Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and sighted listeners," in *Proc. Interspeech*, Chiba, Japan, Sept. 2010, pp. 2186–2189.

[22] M. Barouti, K. Papadopoulos, and G. Kouroupetroglou, "Synthetic and natural speech intelligibility in individuals with visual impairments: Effects of experience and presentation rate," in *European AAATE Conference*, Portugal, 2013, pp. 695–699.

Intelligibility analysis of fast synthesized speech

Cassia Valentini-Botinhao[1], *Markus Toman*[2], *Michael Pucher*[2], *Dietmar Schabus*[2], *Junichi Yamagishi*[1,3]

[1] The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
[2] Telecommunications Research Center Vienna (FTW), Austria
[3] National Institute of Informatics, Japan

{cvbotinh,jyamagis}@inf.ed.ac.uk, {toman,pucher,schabus}@ftw.at

Abstract

In this paper we analyse the effect of speech corpus and compression method on the intelligibility of synthesized speech at fast rates. We recorded English and German language voice talents at a normal and a fast speaking rate and trained an HSMM-based synthesis system based on the normal and the fast data of each speaker. We compared three compression methods: scaling the variance of the state duration model, interpolating the duration models of the fast and the normal voices, and applying a linear compression method to generated speech. Word recognition results for the English voices show that generating speech at normal speaking rate and then applying linear compression resulted in the most intelligible speech at all tested rates. A similar result was found when evaluating the intelligibility of the natural speech corpus. For the German voices, interpolation was found to be better at moderate speaking rates but the linear method was again more successful at very high rates, for both blind and sighted participants. These results indicate that using fast speech data does not necessarily create more intelligible voices and that linear compression can more reliably provide higher intelligibility, particularly at higher rates.

Index Terms: fast speech, HMM-based speech synthesis, blind users

1. Introduction

Blind individuals are capable of understanding speech reproduced at considerably high speaking rates [1]. As screen readers become an essential computer interface for blind users, a challenge arises: how to provide intelligible synthesized speech at such high rates? The standard HSMM-based synthesizer [2] models speech duration by using explicit state duration distributions but for very fast speaking rates this is often not sufficient [3]. It is also unclear whether using fast speech to train a synthesizer can create more intelligible fast synthesized speech than other sorts of compression methods.

Fast speech production and perception has been the target of various studies [4–8]. When producing fast speech vowels are compressed more than consonants [4] and both word-level [5] and sentence-level [6] stressed syllables are compressed less than unstressed ones. Yet another important aspect of fast speech is the significant reduction of pauses. It is claimed that reducing pauses is in fact the strongest acoustic change when speaking faster [7], most probably due to the limitations of how much speakers can speed up their articulation rate [8]. It is argued that these observed changes are the result of an attempt to preserve the aspects of speech that carry more information. The presence of pauses however have been shown to contribute to intelligibility [9].

It has been shown that fast speech (around 1.56 times faster than normal speech) is harder to process, in terms of reaction time, and also preferred less than linearly compressed speech [5, 10]. Linearly compressed speech was found to be more intelligible and better liked than a nonlinearly compressed version of speech where fast speech prosodic patterns were mimicked [5]. The author claims that possibly the only nonlinear aspect of natural fast speech duration changes that can improve intelligibility at high speaking rates is pause removal but only when rates are relatively high [10]. Another nonlinear compression method is the MACH1 algorithm [11]. This method is also based on the acoustics of fast speech with the addition of compressed pauses. It has been shown that at high speaking rates (2.5 and 4.1) MACH1 improves comprehension and is preferable to linearly compressed speech but no advantage was found at the fast speech speaking rate (1.4) [12].

Fast synthesized speech generated by a formant-based system was found to be less intelligible than fast natural speech and the intelligibility gap grows with the speaking rate [13]. More recently the authors in [14] evaluated the intelligibility of a wider range of synthesizers: formant, diphone, unit selection and HMM-based. It was found that the unit selection systems were more intelligible across speech rates. In this evaluation, however, the evaluated synthesizers were based on different speakers and the compression methods adopted by each system were not reported. Literature on fast synthesized speech also focuses on the effect on blind listeners. To improve duration control of HMM-based systems for blind individuals [3] proposed a model interpolation method. Pucher et al. found that interpolating between a model trained with normal and a model trained with fast speech data results in speech that is more intelligible and preferable, for both blind and non blind individuals.

In this paper, we are interested in analysing two aspects of fast synthesized speech. First, the corpus used to train synthesis models, i.e., is it really necessary or even helpful to use fast speech recordings? Second, compression method; which is more effective: a nonlinear manipulation of speech duration or a linear compression method? We evaluate intelligibility of a fast and a normal female Scottish voice and a German male voice, compressed using two nonlinear and one linear method and presented to listeners at different rates.

This paper is organized as follows: Section 2 describes the methods used to create synthetic speech at fast rates, Section 3 presents the corpus used for training the synthesis models and details on how models were trained, Section 4 shows the design and results of intelligibility listening experiments, Section 5 presents a discussion on these results followed by conclusions in Section 6.

2. Compression methods

In this section, we describe methods that can create synthetic speech at fast rates, referred to here as compression methods. The first two methods we describe manipulate the state duration model parameters (mean and/or variance) while the third is applied to the synthesized speech waveform. The first two methods are considered to be nonlinear as each state is compressed at a different rate, as opposed to the third method, which is a linear method that compresses the waveform uniformly across time.

2.1. Variance scaling

Variance scaling is the standard method for duration control in HSMM-based synthesis [15]. With this method we compute the duration of state i as:

$$d_i = \mu_i + \rho\sigma_i \quad (1)$$

where μ_i and σ_i are the mean and variance of the state duration model and ρ is a factor that controls the variance scaling. When $\rho = 0$ the duration is set to the mean state duration, $\rho > 0$ makes synthetic speech slower and $\rho < 0$ faster. The scaling factor is fixed across all states. State duration control is then proportional only to the variance: states whose duration model variance is higher will be compressed more. With this method we can potentially capture certain non-linearities between normal and fast speech durations.

2.2. Model interpolation and extrapolation

In previous work with fast synthetic speech [3], we showed that model interpolation [16, 17] can outperform the variance scaling method in terms of intelligibility and listener preference. Given two voice models of the same speaker trained with speech recorded at normal and fast speaking rates, the most successful method in that study was one that applied interpolation between duration models, using the normal speaking rate models of cepstral, fundamental frequency and aperiodicity features. The interpolated duration d_i for state i is calculated as:

$$d_i = (1-\alpha)\,\mu_i^n + \alpha\,\mu_i^f \quad (2)$$

where μ_i^n and μ_i^f denote the mean duration of state i in the normal and fast duration model and α is the interpolation ratio to control the speaking rate. We can generate speaking rates beyond the rate of the fast model by extrapolating ($\alpha > 1$).

For the experiments in the present paper, we have implemented an additional constraint in this method. It is possible that for a given state of a given phone, the mean duration μ_i^f from the fast model is actually longer than the mean duration μ_i^n of the normal model, causing the speech segments generated for this state to become *slower* with growing α. If this is the case, we do not interpolate or extrapolate, but apply a linear factor β to μ_i^n, where β reflects the overall mean speaking rate difference between the normal and the fast voice models ($\beta = 1/1.55$ in our experiments).

2.3. WSOLA

The waveform similarity overlap and add (WSOLA) method proposed in [18] was chosen here to illustrate the effect of a linear compression. The method provides high enough quality while being computationally efficient and robust [18]. In WSOLA speech frames to be overlapped are first cross-correlated to provide an appropriate time shift that ensures frames are added coherently, inspired by the idea that modified speech should maintain maximum local similarity to the original signal.

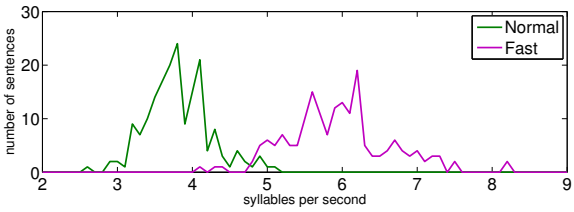


Figure 1: *English TTS voices: syllables per second distribution.*

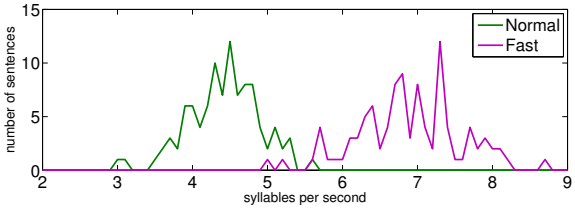


Figure 2: *German TTS voices: syllables per second distribution.*

3. Speech databases and voices

We present the English and German corpora used in our experiments as well as details of how we trained the synthetic voices.

3.1. English – corpus and voices

We recorded a Scottish female voice talent reading 4600 sentences at a normal speed and 800 sentences at a fast speed with the instruction to speak as fast as possible while maintaining intelligibility.

To train the acoustic models, we extracted the following features from the natural speech sampled at 48 kHz: 59 Mel cepstral coefficients [19], Mel scale fundamental frequency F0 and 25 aperiodicity band energies extracted using STRAIGHT [20]. We used a hidden semi-Markov model as the acoustic model. The observation vectors for the spectral and excitation parameters contained static, delta and delta-delta values. One stream was set for the spectrum, three for F0 and one for aperiodicity.

We trained two voices. What we refer to as the model N, is a voice trained only with speech produced at the normal speaking rate. This model was adapted [21] using the 800 sentences of fast speech to create what is referred to as the voice F.

To measure the speaking rate of each synthetic voice we calculated the rate of syllables per second (SPS) and words per minute (WPM) for each sentence used in the evaluation. On average the SPS values of the normal and the fast voice are 3.8 and 6.0 while the values for WPM are 206.7 and 320.9, respectively. Speech synthesized using the fast model is around 1.55 times faster, which agrees with the literature [5] on naturally produced fast speech. Fig. 1 shows the histogram of SPS across synthesized sentence for each voice.

3.2. German – corpus and voices

We used a very similar setup to record and train the German voice. We recorded an Austrian German voice talent reading

W-N	WSOLA applied to normal speech
W-F	WSOLA applied to fast speech
V-N	Variance scaling applied to model N
V-F	Variance scaling applied to model F
I	Interpolation of model N and F

Table 1: *Methods evaluated.*

4387 sentences at a normal and 198 sentences at a fast speaking rate. The German recordings were sampled at 44.1 kHz and we extracted 39 Mel cepstral coefficients. Otherwise the procedure and parameters were the same as for English.

The average SPS values for the normal and fast German synthetic voices are 4.5 and 7.0, and the WPM values are 152.7 and 237.1. The German voice is thus considerably faster than the English voice, at both speaking rates. Interestingly, the fast model is also about 1.55 times faster than the normal model, i.e., the speed-up factor between the two English models and between the two German models is the same. Fig. 2 shows the SPS distribution for the two German models.

4. Evaluation

We conducted two listening experiments with the English voices, one using natural speech and the other TTS; while for the German data only the TTS voices were evaluated, but by both blind and sighted individuals.

We evaluate intelligibility at four different speaking rates: 1.25, fast (the speed of fast speech), 2.0 and 3.0, where numbers refer to speed increase with respect to the normal voice calculated sentence by sentence, remembering here that fast speech is around 1.55 times faster than normal speech. Rates were chosen to reflect conversational, fast and two ultra fast speeds.

The methods we evaluate are presented in Table 1 [1]. Not all methods are evaluated at all speaking rates, for instance at rates smaller or equal to the fast rate W-F, V-F and I were not evaluated. To generate compressed samples using the variance and the interpolation methods it was necessary to progressively change the scale factor to obtain the desired duration. The implementation of WSOLA used here was provided as support material for [22].

Results are presented as percentage of word errors, calculated per listener as the percentage of words that were not transcribed, misspellings taken into account.

4.1. English – evaluation

We evaluate the intelligibility of natural speech compressed only with the WSOLA algorithm as the other two methods can not be applied directly to natural speech. We compare two natural speech compressions: W-N and W-F, compression applied to the normal and the fast speech databases.

For the TTS evaluation, we compare the three different compression methods described in Section 2, although not all methods were evaluated for all speaking rates.

4.1.1. Listening experiment

We performed two listening experiments, one with natural speech and the other with the TTS voices. Each experiment was performed by 20 native English speakers without TTS expertise. Each participant transcribed 10 different sentences for each of the tested methods. The natural speech sentences were selected from news articles while for the TTS experiments sentences were chosen from the first few sets of the Harvard dataset [23].

[1] Speech samples used in the evaluation can be found at: http://wiki.inf.ed.ac.uk/CSTR/SalbProject

4.1.2. Results

Fig. 3 shows the percentage of word errors for each speaking rate obtained in the natural (blue) and TTS (red) experiments.

We can see that the TTS voices created using WSOLA are the most intelligible across all tested speaking rates and that this advantage grows with increasing speaking rate. At the fastest rate the TTS voice W-N results in less than 20 % word errors while the word errors obtained by V-N, V-F and I are higher than 40 %, i.e., errors doubled. Interpolation is slightly better than variance scaling, although not significantly.

Word errors are smaller when compressing speech synthesized from the normal model (W-N) as opposed to a fast model (W-F), as results for speaking rate 2xs show. Although differences are not significant, error levels for by V-F and I are slightly smaller than V-N at all speaking rates. At the fast speaking rate, we can see that the fast voice is less intelligible than the normal voice with linear compression applied.

Compared to the natural speech results (in blue) we can see that error scores are significantly higher for TTS voices. The increase in error seen for W-F compared to W-N for TTS voices can also be observed for natural speech, pointing to the fact that the fast natural speech is also less intelligible than linearly compressed normal speech.

4.2. German – evaluation

A similar evaluation was carried out for German to assess the intelligibility achieved by the methods described in Section 2.

4.2.1. Listening experiment

For the German data, only TTS voices were evaluated. The participants in the listening test consisted of two groups: 16 blind or visually impaired participants, 15 of whom reported using TTS in their everyday life, and 16 sighted participants with no TTS expertise. Each participant transcribed 100 different sentences such that within a participant group, every combination of method and speaking rate was evaluated once. The sentences were selected from news articles and parliamentary speeches.

4.2.2. Results

The results are shown in Fig. 4, where the two bars per condition reflect the results from the two participant groups. As expected, the blind listeners (yellow) generally achieve lower word error percentages than the sighted listeners (red).

Similar to the English results, WSOLA compression of speech synthesized from the normal model (W-N) is the best method overall. However, up to speaking rate 2xs, both WSOLA of fast speech (W-F) and interpolation (I) yield results competitive to W-N. At the "fast" rate, where both W-F and I (and also V-F) are equivalent to simply the fast voice model, these methods even achieve significantly better results than W-N for the sighted listeners. At the "fast" and 2xs rates, W-F and I perform significantly better than variance scaling of the normal model (V-N), confirming the results of [3]. However, we see a very clear advantage of the WSOLA methods at the fastest rate 3xs, where the error percentages of V-N, V-F and I are much higher, yielding a picture similar to the English results at 2xs. There is no significant difference between W-N and W-F at the 2xs and 3xs rates.

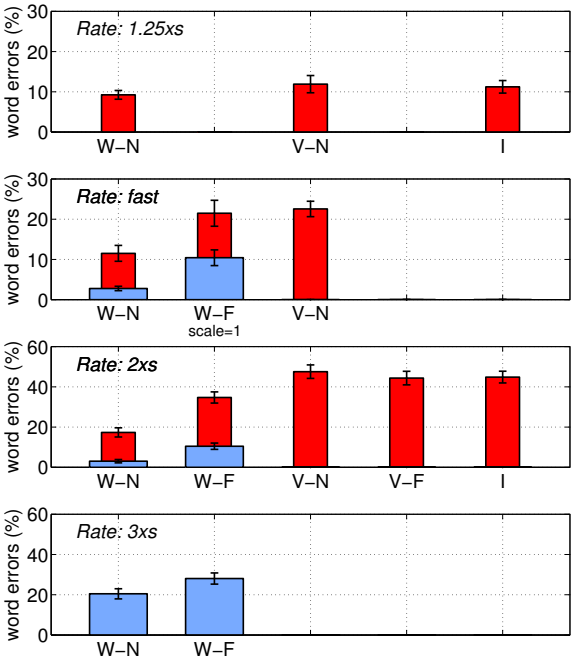


Figure 3: *English results: TTS (red) and natural speech (blue).*

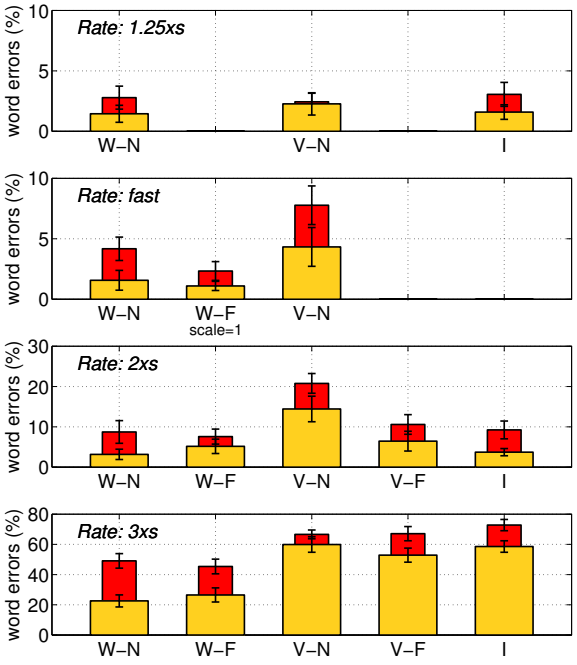


Figure 4: *German results: Sighted (red) and blind/visually impaired listeners (yellow).*

5. Discussion

As found in other studies of natural fast speech [5, 10], our results using the English data also indicate that linear compression can produce more intelligible voices than nonlinear methods based on or directly derived from the acoustics of fast speech. English results show that there is no additional advantage to using recordings of fast speech to build a synthetic voice and it is possible to maintain intelligibility at higher speaking rates by applying a simple linear compression method to the synthesized waveform. This is supported by results with the natural speech corpus, where we also found that fast natural speech is not as intelligible as linearly compressed normal speech.

Results for the German data tell a slightly different story. For German, we also see that linear compression is beneficial at very high speaking rates (3xs) compared to interpolation and variance scaling. For lower speaking rates (2xs), we find that interpolation is equally good as linear compression. This indicates a potential use of a combined method of interpolation for fast speaking rates and linear compression for ultra-fast speaking rates. We hypothesize that different results were found for the German data due to the inherent higher intelligibility of the German fast speech, which can also be seen in the performance differences of linear compression of synthesized speech from fast models (W-F) which performs better for the German data. We want to investigate this hypothesis in the future by carrying out a detailed analysis of fast speech durations from different speakers. Concerning the performance of blind listeners we can confirm results presented in previous studies [1, 3], which show that blind listeners achieve lower word-error-rates than non-blind listeners.

Considering results on both databases we hypothesize that methods that use recordings of fast speech such as adaptation or interpolation are perhaps only as intelligible as the fast data they use. Relying on having fast speech that is intelligible enough is challenging as this data is quite difficult to produce considering that both of our speakers are voice talents. Using more recordings of fast speech is also not helpful as more fast sentences were used for the English voices. Moreover it is not yet clear how to reach very high speaking rates with model interpolation and adaptation as these methods are limited by the fact that no skip is allowed. The weak performance of the variance scaling method for fast speaking rates (2xs, 3xs) is in agreement with the poor results obtained by HMM-based voices in [14].

6. Conclusion

We showed that linear compression outperforms the variance scaling and interpolation methods for ultra-fast (3xs) speaking rates in German and English. For fast speaking rates (2xs) linear compression outperformed other methods for English while being as good as interpolation for German. In general we see that the usage of fast speech data in interpolation (I) or linear compression (W-F) is dependent on the quality of the data.

As future work, we plan to evaluate the intelligibility of the German language corpus and the TTS voices in English with blind participants as well. Additionally, we plan to analyse the acoustic properties of both fast speech corpus in more detail in order to explain the differences in their intelligibility.

7. Acknowledgement

This work was supported by the BMWF - Sparkling Science project *Sprachsynthese von Auditiven Lehrbüchern für Blinde SchülerInnen* (SALB). The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH is funded within the program COMET - Competence Centers for Excellent Technologies by BMVIT, BMWA, and the City of Vienna. The COMET program is managed by the FFG.

8. References

[1] A. Moos and J. Trouvain, "Comprehension of ultra-fast speech – blind vs. 'normally hearing' persons," in *Proc. Int. Congress of Phonetic Sciences*, vol. 1, 2007, pp. 677–680.

[2] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-markov model-based speech synthesis system," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 825–834, 2007.

[3] M. Pucher, D. Schabus, and J. Yamagishi, "Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and sighted listeners," in *Proc. Interspeech*, Chiba, Japan, Sept. 2010, pp. 2186–2189.

[4] T. Gay, "Effect of speaking rate on vowel formant movements," *J. Acoust. Soc. Am.*, vol. 63, no. 1, pp. 223–230, 1978.

[5] E. Janse, S. Nooteboom, and H. Quené, "Word-level intelligibility of time-compressed speech: Prosodic and segmental factors," *Speech Comm.*, vol. 41, no. 2, pp. 287–301, 2003.

[6] R. F. Port, "Linguistic timing factors in combination," *J. Acoust. Soc. Am.*, vol. 69, no. 1, pp. 262–274, 1981.

[7] F. Goldman-Eisler, *Psycholinguistics: Experiments in spontaneous speech*. London: Academic Press, 1968.

[8] R. Greisbach, "Reading aloud at maximal speed," *Speech Comm.*, vol. 11, no. 4-5, pp. 469 – 473, 1992.

[9] A. A. Sanderman and R. Collier, "Prosodic phrasing and comprehension," *Language and Speech*, vol. 40, no. 4, pp. 391–409, 1997.

[10] E. Janse, "Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech," *Speech Comm.*, vol. 42, no. 2, pp. 155–173, 2004.

[11] M. Covell, M. Withgott, and M. Slaney, "Mach1: Nonuniform time-scale modification of speech," in *Proc. ICASSP*, vol. 1. Seattle, USA: IEEE, May 1998, pp. 349–352.

[12] L. He and A. Gupta, "Exploring benefits of non-linear time compression," in *Proc. ACM Int. Conf. on Multimedia*. Ottawa, Canada: ACM, Sept. 2001, pp. 382–391.

[13] J. Lebeter and S. Saunders, "The effects of time compression on the comprehension of natural and synthetic speech," *Working Papers of the Linguistics Circle*, vol. 20, no. 1, pp. 63–81, 2010.

[14] A. K. Syrdal, H. T. Bunnell, S. R. Hertz, T. Mishra, M. F. Spiegel, C. Bickley, D. Rekart, and M. J. Makashay, "Text-to-speech intelligibility across speech rates." in *Proc. Interspeech*, Portland, USA, Sept. 2012.

[15] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," in *Proc. ICSLP*, Sydney, Australia, Dec. 1998, pp. 29–32.

[16] ——, "Speaker interpolation for HMM-based speech synthesis system," *Acoustical Science and Technology*, vol. 21, no. 4, pp. 199–206, Jan. 2000.

[17] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE Trans. Inf. Syst.*, vol. E88-D, no. 11, pp. 2484–2491, Nov. 2005.

[18] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. ICASSP*, vol. 2, April 1993, pp. 554–557.

[19] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, vol. 1, San Francisco, USA, March 1992, pp. 137–140.

[20] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Comm.*, vol. 27, pp. 187–207, 1999.

[21] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 66 –83, 2009.

[22] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2010.

[23] IEEE, "IEEE recommended practice for speech quality measurement," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225 – 246, 1969.

Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and sighted listeners

Michael Pucher[1], *Dietmar Schabus*[1], *Junichi Yamagishi*[2]

[1]Telecommunications Research Center Vienna (FTW), Austria
[2]The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
pucher@ftw.at, schabus@ftw.at, jyamagis@inf.ed.ac.uk

Abstract

In this paper we evaluate a method for generating synthetic speech at high speaking rates based on the interpolation of hidden semi-Markov models (HSMMs) trained on speech data recorded at normal and fast speaking rates. The subjective evaluation was carried out with both blind listeners, who are used to very fast speaking rates, and sighted listeners. We show that we can achieve a better intelligibility rate and higher voice quality with this method compared to standard HSMM-based duration modeling. We also evaluate duration modeling with the interpolation of all the acoustic features including not only duration but also spectral and F0 models. An analysis of the mean squared error (MSE) of standard HSMM-based duration modeling for fast speech identifies problematic linguistic contexts for duration modeling.
Index Terms: speech synthesis, fast speech, hidden semi-Markov model

1. Introduction

It is well known that synthetic speech at very high speaking rates is frequently used by blind users to increase the amount of presented information. In data-driven approaches, however, this may lead to the severe degradation of synthetic speech quality, especially at very fast speaking rates. The standard HSMM-based duration modeling [1] is already able to model certain non-linearities between normal and fast speech units since it uses explicit state duration distributions and can thereby take the duration variance of units into account, but for very fast speaking rates this is not sufficient. We therefore propose a duration control method using a model interpolation technique, where we can continuously interpolate HSMMs for normal and fast speaking rate. The HSMMs for fast speaking rate are adapted from HSMMs for the normal speaking rate. In addition to the interpolation between the normal and fast speaking rate, we can also use the extrapolation between them to achieve very fast speaking rates that go beyond the recorded original speaking rates. A conventional study [2] already showed that an HMM-based synthesizer with interpolated duration models can outperform a synthesizer with rule based duration model. Their models were, however, based on so called Hayashi's quantification method I and were theoretically different from our methods based on HSMM interpolation and adaptation techniques, which are available from the HTS toolkit today [3].

Some studies shown that the complex duration changes between normal and fast speech are present at several linguistic levels [4] at the same time. Therefore we employ context-dependent linear regression functions for the HSMM duration adaptation to achieve the duration changes at different linguistic levels. Our contexts used include high-level linguistic features such as syllable information, phrase information etc. The use of HSMM duration adaptation has another advantages. For example this makes online process of the proposed duration control technique possible since normal and fast duration models have the same tying structure and we can straightforwardly perform the interpolation online. This also makes the analysis of the modeling error of standard duration modeling for fast durations easier.

Table 1: Three duration modeling methods used in our evaluation.

Method	Description
SPO–SPO+F	Interpolation between SPO voice and SPO voice with fast duration model.
SPO–SPF	Interpolation between SPO voice and SPO voice with fast duration, spectrum, and F0 model.
SPO	HMM-based duration modeling using acceleration coefficient ρ.

For the evaluation we carried out a comprehension and pairwise comparison test with both blind and sighted listeners. We confirmed that both groups of the listeners preferred sentences generated with our method than the conventional method. The proposed method could also achieve lower word error rates (WER) in the comprehension test. The blind listeners were especially good in understanding sentences at fast speaking rates (6-9 syllables per second) compared to non-blind listeners.

2. HSMM Duration Modeling

2.1. Duration modeling methods

All synthetic voices used are built using the framework of a speaker adaptive HMM-based speech synthesis system. Detailed description of the system is given in [5]. Note that our model adaptation is a two-step approach: the first adaptation is for speaker transformation and the second adaptation is for speaking rate adaptation. First we trained an average voice model using several background speakers from speech data at normal speaking rate [6]. We then adapted the average voice model to two Austrian German male speakers (SPO, HPO) using speech data having the normal speaking rates. Further we transformed the adapted speaker-specific models by using fast speech uttered by the same speakers. We call the adapted models for the fast speaking rates SPF and HPF, respectively. As adaptation data we used a phonetically balanced corpus con-

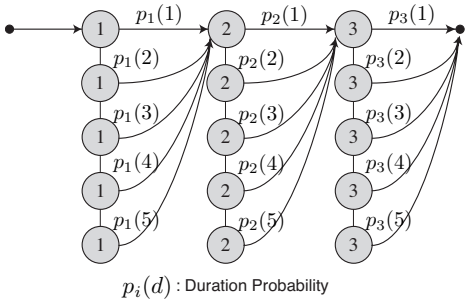


Figure 1: WFST-like illustration of duration models used for TTS systems. Duration probabilities p_i are also transformed to target speakers during speaker adaptation.

sists of ca. 300 sentences for each of the normal speaking rate and fast speaking rate uttered by each speaker. Table 1 shows three methods that were used in the evaluation. SPO–SPO+F and SPO–SPF are proposed methods that use the interpolations of adapted HSMMs.

To make the differences between the three methods clearer, we explain the temporal structure of the HSMMs [8] and its adaptation. In addition to observations such as the mel-cepstrum and fundamental frequency, each semi-Markov state has a stack of states with associated duration probabilities p_i, illustrated in Fig. 1. The duration probabilities p_i are characterized by Gaussian pdfs, and the mean and variance of the pdfs

$$p_i(d) = \mathcal{N}(d; \mu_i, \sigma_i^2). \tag{1}$$

In the HSMM-based parameter generation [1], we use the mean sequence $(\mu_1, \cdots, \mu_N)$ of the Gaussian pdfs corresponding to a given input unit sequence as the most likely sequence. Here N represents the number of states. The easiest and simplest way to control duration is to manipulate the mean of each state using the variance of the state

$$\hat{\mu}_i = \mu_i + \rho\sigma_i^2. \tag{2}$$

and to use a sequence $(\hat{\mu}_1, \cdots, \hat{\mu}_N)$ as a state sequence for the parameter generation. Here ρ is an acceleration coefficient and $\rho > 0$ makes synthetic speech slower and $\rho < 0$ makes synthetic speech faster.

Another way is to transform model parameters for the Gaussian pdfs using a small amount of data for fast speech. There are several possible ways for the transformation and here we employ the CMLLR transform [5], which is given by

$$\mu_i^{\text{fast}} = A_i\mu_i + B_i, \tag{3}$$
$$\sigma_i^{2\ \text{fast}} = A_i^2\sigma_i^2. \tag{4}$$

Here linear regression coefficients A_i and B_i are context-dependent and they are tied through context decision trees having a lot of linguistic questions. To produce speech at various speaking rates, the adapted mean vectors are further interpolated by the original mean vectors

$$\tilde{\mu}_i = (1-w)\mu_i + w\mu_i^{\text{fast}} \tag{5}$$
$$= (1 - w + wA_i)\mu_i + wB_i, \tag{6}$$

where w is the interpolation ratio to control the speaking rate. This interpolation is performed along the state-dependent linear functions obtained from the normal and fast speech. Then a

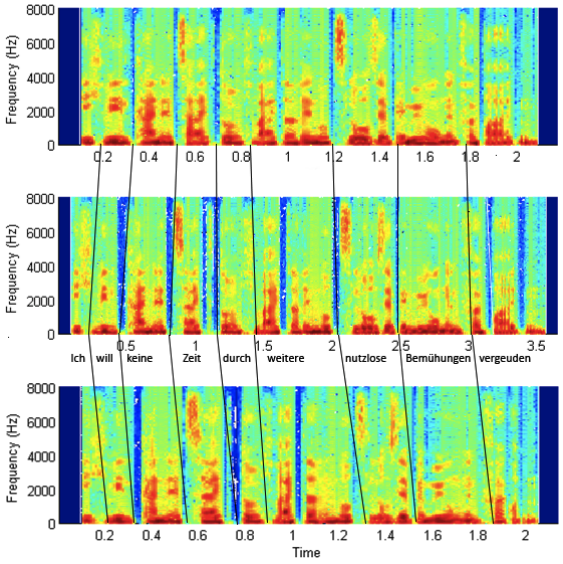


Figure 2: Spectrum of fast sentence with duration interpolation (SPO–SPO+F, top), normal duration (SPO, $\rho = 0$, middle), and fast duration (SPO, $\rho < 0$, bottom).

sequence $(\tilde{\mu}_1, \cdots, \tilde{\mu}_N)$ is used as a state sequence for the parameter generation. The same idea may be use other acoustic features and system SPO–SPF uses this idea for all the features (spectrum, F0, and duration). System SPO–SPO+F uses this interpolation only for duration pdfs.

Figure 2 shows spectra for the utterance "Ich will keine Zeit durch weitere nutzlose Bemühungen vergeuden (I do not want to spend time on additional useless efforts.)". Especially the last word is squeezed with standard duration modeling of fast speech (SPO), which makes it hardly audible. With interpolation this word is much better modeled (top image). Through interpolation we can achieve a better non-linear modeling of duration since we take into account the duration changes from normal to fast speech for contextually modeled units.

2.2. Comparison of adapted duration models

It is important for us to analyze how different the duration values produced by using (2) are from duration using (6). For such analysis, the acceleration coefficient ρ that is necessary to change from duration μ_i (normal duration) to duration μ_i^{fast} (fast duration) may be calculated by

$$\rho = \frac{\mu_i^{\text{fast}} - \mu_i}{\sigma_i^2}. \tag{7}$$

Using (7) , we can define the mean-squared-error between normal duration and fast duration models that would be produced by using this acceleration coefficient for all models (leaf nodes in the duration clustering tree) as follows:

$$e_i = \frac{1}{M}\sum_{i=1}^{M}((\mu_i + \rho\sigma_i^2) - \mu_i^{\text{fast}})^2. \tag{8}$$

This value tells us something about the duration errors that a model produces and thereby about the quality that we have

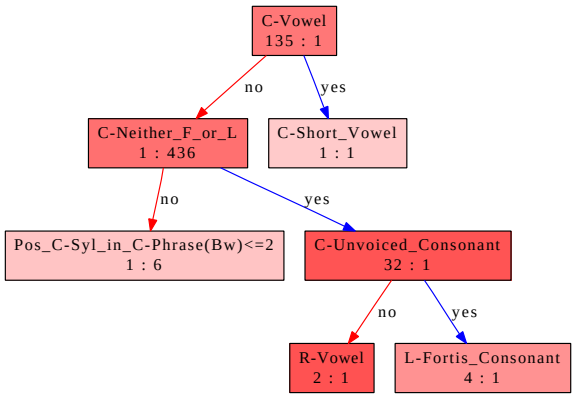


Figure 3: Top of the duration clustering tree. Nodes filled in a darker shade of red have greater average error defined by Eq. (9).

achieved in modeling that specific context. Furthermore we define the error for each non-terminal node n as the average error of all leaves under n:

$$\bar{e}_n = \frac{\sum_{k \in \text{leaves}(n)} e_k}{|\text{leaves}(n)|}. \quad (9)$$

Figure 3 shows the top of the duration clustering tree with the nodes colored according to their error (Equation 9) for speaker SPO. Looking at the entire tree (which has 1897 leaves and hence 1896 inner nodes) reveals that the subtree rooted at the node labelled "R-Vowel" has particularly many problematic models. Each node in the figure is labelled with the corresponding question as well as the ratio of the errors of its children, $e_{i \to no} : e_{i \to yes}$. For example, the root node question asks whether the central phone is a vowel. We see that the average error for non-vowels is 135 times as big as for vowels. Among the non-vowels, phones which belong to the class "neither fortis or lenis" are particularly error-prone, and of these, unvoiced consonants are not quite as bad. During the construction of the tree, the class "Neither_F_or_L" was defined as containing the phones /l/ , /m/ , /n/ , /N/ and /h/ . Of these, only /h/ belongs to the unvoiced consonants, hence the central phone for all models under the node labelled "R-vowel" must be one of /l/ , /m/ , /n/ , /N/. We can see how bad this subtree really is by looking at the cumulative error made by all its leaves (without averaging): The subtree rooted at "R-Vowel" accounts for more than 98% of the total error in the whole tree, but it only accounts for about 13% of the number of leaves.

It is known that the duration of consonants in fast speech is more problematic than that of vowels [REFERENCE], but why exactly those four stand out so clearly should be further investigated in future work. For speaker HPO, we do not see such clearly distinguished subtrees, however the general trend of consonants having greater average error is confirmed also here. This could be due to more inconsistency of speaker HPO in terms of duration.

3. Evaluation

We evaluated two different male speaker's voices namely SPO and HPO. The duration modeling methods for one voice are described in Table 1. We generated utterances with 7 different durations using the standard HSMM-based synthesis duration method (SPO), interpolation between normal and fast duration model (SPO–SPO+F), and interpolation between normal and fast duration, spectrum, and F0 model (SPO–SPF). In the pairwise comparison we compare the same utterance with the same duration and speaker using different modeling methods.

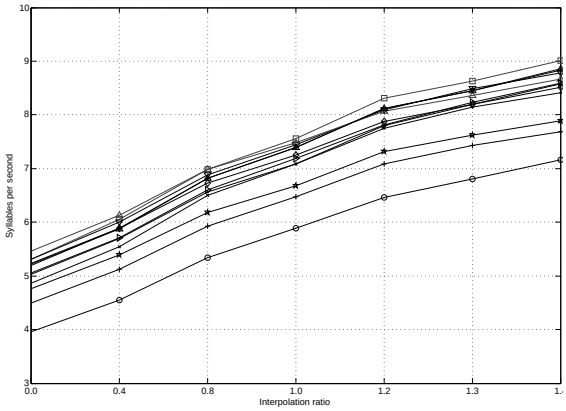


Figure 4: Syllables per second for SPO sentences.

Table 2: Overall word-error rate (WER) and sentence-error-rate (SER) for blind and sighted users.

Users	WER / SER in %	# sent.
Blind users	19.5 / 57.4	108
Sighted users	24.4 / 70.4	216

For the evaluation we had 18 sighted listeners (24 to 55 years; 9 female / 9 male) and 9 blind listeners (28 to 56 years; 4 female / 5 male). Within the group of blind listeners we had 2 visually impaired listeners. The users first had to listen once to 12 sentences and write down what they heard (comprehension test). Afterwards they had to listen to pairs of sentences and decide which sentence they prefer in terms of overall quality. In the pairwise comparison each pair was listened to at least two times by some user.

3.1. Duration of prompts

The duration of prompts is determined by the interpolation ratio and therefore depends on the speaker's duration model. As interpolation ratio we used the following values $[0.0, 0.4, 0.8, 1.0, 1.2, 1.3, 1.4]$. With 0.0 and 1.0 no interpolation is done and only the normal or fast duration model is used. $[1.2, 1.3, 1.4]$ are extrapolation ratios to achieve very fast speaking rates. For the evaluation we had 12 different prompts. Figure 4 shows how many syllables per second are realized for the different sentences by speaker SPO. The fastest sentences contain up to 9 syllables per second.

3.2. Comprehension

Table 2 shows the error rates for blind and sighted users. Blind listeners are better in understanding fast speech than sighted listeners. This can also be seen from Figure 5 where we plotted the word-error-rates for the different interpolation ratios (i.e. durations). One can see that blind users are especially good at recognizing fast speech $[1.3, 1.4]$ where the error rate for sighted users is much higher. While the WERs for sighted listeners are almost monotonically increasing, WERs for blind listeners are

Table 3: Word-error rate (WER) and sentence-error-rate (SER) per method.

Method	WER / SER in %	# sent.
SPO–SPO+F	16.4 / 61.1	54
SPO–SPF	21.1 / 66.7	54
SPO	23.2 / 66.7	54
HPO–HPO+F	24.3 / 66.7	54
HPO–HPF	25.3 / 63.0	54
HPO	26.3 / 72.2	54

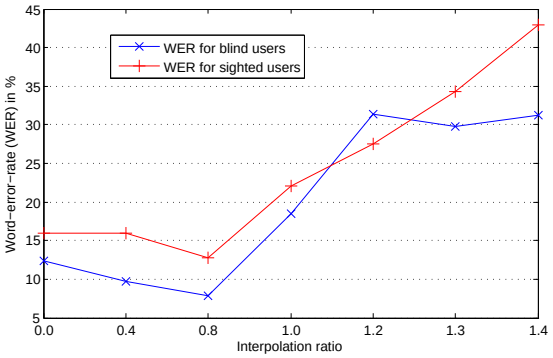


Figure 5: Word-error-rate for blind and sighted users per interpolation ratio (duration).

flat from 1.2 to 1.4.

Other work [7] has shown that the error rate for non-blind users increases fast from a rate of 10.5 syllables per seconds onwards. In that study, comprehension was subjectively measured by asking listeners how much they could understand of the text. As shown by Figure 4 and 5, the division between blind and sighted users concerning understanding can already be seen at around 8 syllables per second when using the objective word error rate measure.

3.3. Pairwise comparison

Figure 6 shows the preference rates for the different methods for all listeners. We see that the adaptive interpolation method where just the duration model is interpolated outperforms the other methods. SPO–SPO+F and HPO–HPO+F are significantly different from the other two methods ($p < 0.05$). The difference is smaller for the HPO voices, since these voices are of a general lower quality than the SPO voices. This lower quality makes the subtle differences of duration modeling more difficult to perceive.

4. Conclusion and future work

We have presented a HSMM-based method for the synthesis of fast speech that outperforms other standard methods on understanding and overall quality. Especially for blind users it is important to have high quality synthesis techniques for fast speech. The adaptive method that we presented can be used with limited amounts of fast adaptation data.

In future work we want to investigate the duration rates that are used by blind users of speech synthesis in their everyday use. Furthermore we want to analyze the error of duration modeling on the basis of corpora not only with a comparison of complete models. We also want to investigate the use of fast speech background models and the use of fast duration models from one speaker for another speaker.

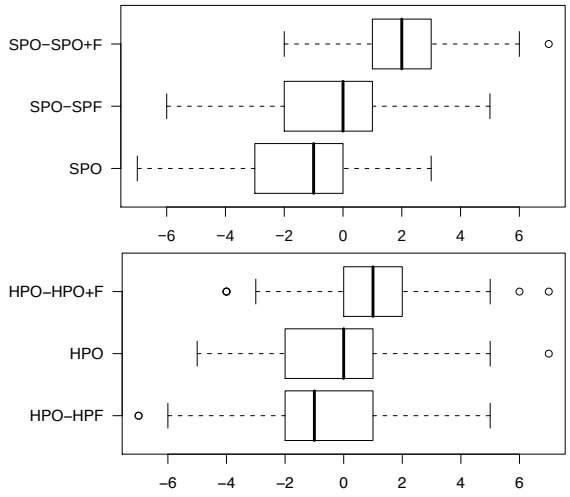


Figure 6: Overall pairwise comparison results per method for speaker SPO (top) and HPO (bottom).

5. Acknowledgements

This work was partly funded by the Vienna Science and Technology Fund (WWTF). The Telecommunications Research Center Vienna (FTW) is supported by the Austrian Government and the City of Vienna within the competence center program COMET. Junichi Yamagishi is funded by EPSRC and EC FP7 collaborative projects (EMIME and LISTA).

6. References

[1] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," IEICE Trans. Inf. & Syst., E90-D, 5, pp.825–834, May 2007

[2] K. Iwano, M. Yamada, T. Togawa, and S. Furui, "Speech-rate-variable HMM-based Japanese TTS system", in Proc. TTS2002, Santa Monica, USA, 2002.

[3] H. Zen, K. Oura, T. Nose, J. Yamagishi, S. Sako, T. Toda, T. Masuko, A.W. Black, and K. Tokuda, "Recent development of the HMM-based speech synthesis system (HTS)," Proc. 2009 Asia-Pacific Signal and Information Processing Association (APSIPA), pp.121–130, Oct. 2009.

[4] E. Janse, "Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech", Speech Communication 42:155–173, 2004.

[5] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis", IEEE Trans. Speech Audio Lang. Process. 17 -(6):12081230 Aug 2009.

[6] M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, V. Strom, "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis" Speech Communication, 52(2):164-179, 2010.

[7] J. Trouvain, "On the comprehension of extremely fast synthetic speech" Saarland working papers in linguistics (SWPL), 1:5-13, 2007.

[8] S. Z. Yu, "Hidden semi-Markov models", Artificial Intelligence, 174(2):215-243, 2010.

2.3 Audio-Visual Text-to-Speech Synthesis

2.3.1 Joint audio-visual modeling

Talking computer-animated characters are now commonplace in entertainment productions such as video games and animated movies. And with the advent of speaking personal assistants, virtual agents will become increasingly important as well. Regardless of the application, speaking characters require lip motion synchronization to recorded or synthetic speech. The quality of this synchronization is important to increase immersion in entertainment products while also being critical to the believability of virtual agents.

State of the art animation for films and also video games is either done entirely by hand or by employing expensive motion capturing technology, which requires extensive manual clean up. Both methods deliver high quality results but are extremely time consuming and expensive. In particular, the amount of dialogue in games has been increasing over the last few years, creating a need for automatic facial animation methods. Likewise, talking virtual agents in dialogue systems need to automatically synchronize their lip motion to synthesized speech to be able to deliver a believable interaction with the user.

However, speech animation is a complex interdisciplinary problem that can be divided into two separate tasks; creating realistic speech dynamics, the rhythm and timing of the articulators and creating realistic deformations on the 3D model, retargeting the dynamics to a particular face. Motion capturing is primarily a means of recording realistic speech dynamics but the retargeting of the recorded motions to specific deformations on a 3D model is a separate problem in the computer graphics field and is out of scope for this work.

While motion capturing of natural speech can accurately capture speech dynamics, no dynamic information is available when using synthesized speech in a dialogue system. Therefore such systems usually rely only on phonetic information. Our work is primarily concerned with creating realistic speech dynamics for synthesized speech. In detail we address the problem of *audiovisual* text-to-speech synthesis (TTS), which is the synthesis of both an acoustic speech signal (TTS in the classical sense), as well as matching visual speech motion parameters given some unseen text as input.

The field of visual speech synthesis is fairly well established and a variety of approaches have been developed since the first rule based systems Cohen and Massaro [1993]. Video-based systems Bregler et al. [1997]; Ezzat et al. [2002] and other data-driven approaches Bailly et al. [2003]; Deng and Neumann [2006]; Theobald et al. [2004] have been developed.

The HMM-based visual speech synthesis systems that have been developed can be broadly categorized into two types: Image-based systems on the one hand use fea-

tures derived directly from the video frames Wang et al. [2011]; Sako et al. [2000] where the resulting synthesis is supposed to look like a video of a real person. Motion capture based approaches Masuko et al. [1998]; Tamura et al. [1998a]; Hofer and Richmond [2010]; Govokhina et al. [2007]; Bailly et al. [2009] on the other hand derive their features from individual facial feature points tracked over time. The advantage of these types of features is that the synthesized motion trajectories can be used to drive any 3D face. Our system is based on motion-capture data but the HSMM-based approach is flexible enough to allow for the synthesis of any type of parameter sequence. Note that our goal is to synthesize both audio speech and motion parameters directly from text, but the models we train can also be used in the less general manner of Wang et al. [2011] and Hofer et al. [2008] (audio unit alignment on speech input followed by visual synthesis, using audiovisual models).

Combining the auditory and visual modalities in one framework requires a synchronous corpus of parametrized facial motion data and acoustic speech data. We have demonstrated in previous work how to build such a corpus [Schabus et al., 2012] and that it is feasible to produce both acoustic speech parameters and animation parameters [Schabus et al., 2011] by a maximum likelihood parameter generation algorithm [Tokuda et al., 2000a] from models that were trained on such a synchronous corpus. In [Schabus et al., 2012] we showed how to generate visual parameters using a speaker-adaptive approach. Here we describe a joint audiovisual speaker-dependent HSMM-based approach for generating visual and acoustic features for different speakers. In statistical data-driven audiovisual synthesis, commonly separate acoustic and visual models are trained [Wang et al., 2011; Sako et al., 2000; Masuko et al., 1998; Tamura et al., 1998a; Hofer and Richmond, 2010; Hofer et al., 2008], sometimes together with an additional explicit time difference model to correctly synchronize the two modalities [Govokhina et al., 2007; Bailly et al., 2009]. In contrast, we propose to train one joint audiovisual model (with acoustic and visual streams), such that the likelihood of the model generating the training data is maximized globally, across the two modalities, during model parameter estimation. This results in a single duration model used for both modalities, thus eliminating the need for additional synchronization measures. In this way, we intend to create simple and direct models for audiovisual speech synthesis, which can cope with most effects of co-articulation and inter-modal asynchrony naturally through five-state quin-phone full-context modeling. Bailly et al. [2009] also argues that states can capture some inter-modal asynchrony since transient and stable parts of the trajectories of different modalities need not necessarily be modeled by the same state, and that multi-phone context models can capture co-articulation effects. Notably, an early work on audiovisual HMM-synthesis Tamura et al. [1999] also applied joint modeling in our sense, however without investigating its benefits in detail. Also, the current HMM-modeling techniques and high-fidelity visual parameter acquisition we use distinguish our work from Tamura et al. [1999].

Therefore the main purpose of our work is to investigate whether the proposed joint audiovisual modeling approach provides clear improvements over separate audio and

visual modeling. We argue that the main weakness of separate modeling stems from the difficulty to capture (and even define) clear temporal unit borders for the visual modality. Our analysis shows that visual-only training yields models which fail to find suitable borders for some phones when we carry out forced alignment on our training data. An explicit audio/video lag model used for modality synchronization, which is trained on such borders (as in Govokhina et al. [2007]; Bailly et al. [2009]) might still suffer from these problems, even if the borders in the training data are hand-labeled (as in Terry [2011]). Furthermore, the quality of the synthesized trajectories themselves can be expected to degrade if observation assignment to units is unclear during training.

On the other hand, there are situations where the targets to which speech needs to be synchronized are much clearer, like singing synthesis Saino et al. [2006], where explicit lag models have been used successfully for synchronizing speech to sheet music (in that case, the sheet music defines fixed and exact synchronization target points in time).

We furthermore consider the description of the system we have built an important part of this work. This system is based on a state-of-the-art HSMM modeling framework and we use current animation-industry-standard motion tracking and character animation technology for the visual modality. In this regard our work differs strongly from conceptually related previous work Sako et al. [2000]; Masuko et al. [1998]; Tamura et al. [1998a].

2.3.2 Adaptive audio-visual modeling

However, as with all HMM-based approaches, large amounts of training data are required to build high quality systems and recording large amounts of video data is even more costly than recording audio data. To address this shortcoming for speakers where limited amounts of data are available, a very successful speaker-adaptive approach has been developed Yamagishi and Kobayashi [2007b]; Yamagishi et al. [2009d] for acoustic HMM-based speech synthesis. A (possibly large) speech database containing multiple speakers is used to train an average voice, where a speaker-adaptive training (SAT) algorithm provides speaker normalization. Then, a voice for a new target speaker can be created by transforming the models of the average voice via speaker adaptation, using (a possibly small amount) of speech data from the target speaker. This allows the creation of many speaker's synthetic voices without requiring large amounts of speech data from each of them. It can be shown that synthetic speech from voice models created in this way is perceived as more natural sounding than synthetic speech from speaker-dependent voice models using the same (target speaker) data Yamagishi and Kobayashi [2007b]. This holds especially for the case where this data is of small amount. The goal of our work is to demonstrate how this speaker-adaptive training approach can be applied to visual speech synthesis.

2.3.3 Visual control of acoustic speech

One key strength of the HSMM-based speech synthesis framework Zen et al. [2007a] lies in its greater flexibility in comparison to waveform concatenation methods, often accredited to the possibility to use model adaptation Yamagishi et al. [2009c] and interpolation Yoshimura et al. [1997]. In addition to these data-driven approaches to diversify the characteristics of synthetic speech, methods that allow more direct control using phonetic background knowledge have been proposed more recently. Acoustic speech characteristics have been successfully modified by exercising control on articulartory Ling et al. [2008] as well as on formant Lei et al. [2011] parameters. This is achieved by training piecewise linear transformations from the models for the articulatory or formant domain to the models for the acoustic domain, using a multimodal data corpus. Similar to these works, in our work we investigate the possibility of using visual speech features based on facial marker motion data to modify and control acoustic synthetic speech. Our work is similar to Ling et al. [2008], but uses more restricted features (e.g., no tongue positions) which are easier to record.

Possible use cases of this include more intuitive control of speech synthesis, the possibility to use physically intuitive data to constrain trajectories as well as the possibility to use this information in language learning to provide clues of required changes.

To investigate the possibility of visual control, we modified the system used in Lei et al. [2011] to control acoustic speech synthesis by visual features (instead of formants). The same line spectral pairs features as in Ling et al. [2008] are adopted as acoustic features in our approach.

To simplify the control from having to modify points in 6 dimensional space, we apply PCA. To modify a given model, the means are transformed into PCA space, modifications are performed relative to the resulting PCA feature vector, and the modified vector is projected back into the original space. No dimensionality reduction is used in this scheme. Also, the trajectory is not modified directly (e.g., by adding to the trajectory values), but the means are changed, thus changing the generated trajectory. This ensures smooth trajectories for the duration of the modified phone and especially at the phone boundaries. As a side effect of the smoothing, the extent of modification is slightly decreased, thus a larger change in the control parameters is required to achieve sufficiently strong effects.

The first PCA component roughly corresponds to mouth opening, while the second and third component can be interpreted as modeling rounding.

We did not carry out a formal evaluation of the effects of the control on the visual speech motion, but synthesis of the entire 37 marker positions can be performed at the loss of some accuracy by calculating a linear regression from the 6 visual control parameters to the full visual parameter space. From examples we looked at during development, it appeared that visual synthesis is still feasible using only these parameters. There is

also some loss of acoustic quality due to the simple transformation and the incomplete explanation of acoustic features by the visual features.

In the sentence *"Ich habe 'bomo' gehört"* (I heard 'bomo'), the two vowels of the nonsense word 'bomo' were modified visually by increasing and decreasing the mean of the first PCA component, corresponding to increased and decreased mouth opening, respectively. The bottom part of the figure shows the effect on the distance between the upper lip and the lower lip markers. The middle part shows the resulting spectrograms for the time segment indicated by vertical lines. The top part of the figure shows the resulting facial marker configurations at the time points indicated by small circles. Compared to the unmodified sample (± 0, center spectrogram, first formant at 287 Hz), the samples with the decreased (-1.5, left spectrogram, first formant at 408 Hz) and increased ($+1.5$, right spectrogram, first formant at 605 Hz) mouth opening exhibit a clearly visible change both in the visual trajectories as well as in the spectral power distributions.

336 IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 8, NO. 2, APRIL 2014

Joint Audiovisual Hidden Semi-Markov Model-Based Speech Synthesis

Dietmar Schabus, Michael Pucher, and Gregor Hofer

***Abstract*—This paper investigates joint speaker-dependent audiovisual Hidden Semi-Markov Models (HSMM) where the visual models produce a sequence of 3D motion tracking data that is used to animate a talking head and the acoustic models are used for speech synthesis. Different acoustic, visual, and joint audiovisual models for four different Austrian German speakers were trained and we show that the joint models perform better compared to other approaches in terms of synchronization quality of the synthesized visual speech. In addition, a detailed analysis of the acoustic and visual alignment is provided for the different models. Importantly, the joint audiovisual modeling does not decrease the acoustic synthetic speech quality compared to acoustic-only modeling so that there is a clear advantage in the common duration model of the joint audiovisual modeling approach that is used for synchronizing acoustic and visual parameter sequences. Finally, it provides a model that integrates the visual and acoustic speech dynamics.**

***Index Terms*—Audiovisual speech synthesis, facial animation, hidden Markov model, HMM-based speech synthesis, speech synthesis, talking head.**

I. Introduction

TALKING computer-animated characters are now commonplace in entertainment productions such as video games and animated movies. And with the advent of speaking personal assistants, virtual agents will become increasingly important as well. Regardless of the application, speaking characters require lip motion synchronization to recorded or synthetic speech. The quality of this synchronization is important to increase immersion in entertainment products while also being critical to the believability of virtual agents.

State of the art animation for films and also video games is either done entirely by hand or by employing expensive motion capturing technology, which requires extensive manual clean up. Both methods deliver high quality results but are extremely time consuming and expensive. In particular, the amount of dialogue in games has been increasing over the last few years, creating a need for automatic facial animation methods. Likewise, talking virtual agents in dialogue systems need to automatically synchronize their lip motion to synthesized speech to be able to deliver a believable interaction with the user.

However, speech animation is a complex interdisciplinary problem that can be divided into two separate tasks; creating realistic speech dynamics, the rhythm and timing of the articulators and creating realistic deformations on the 3D model, retargeting the dynamics to a particular face. Motion capturing is primarily a means of recording realistic speech dynamics but the retargeting of the recorded motions to specific deformations on a 3D model is a separate problem in the computer graphics field and is out of scope for this paper.

While motion capturing of natural speech can accurately capture speech dynamics, no dynamic information is available when using synthesized speech in a dialogue system. Therefore such systems usually rely only on phonetic information. This paper is primarily concerned with creating realistic speech dynamics for synthesized speech. In detail we address the problem of *audiovisual* text-to-speech synthesis (TTS), which is the synthesis of both an acoustic speech signal (TTS in the classical sense), as well as matching visual speech motion parameters given some unseen text as input.

The field of visual speech synthesis is fairly well established and a variety of approaches have been developed since the first rule based systems [1]. Video-based systems [2], [3] and other data-driven approaches [4]–[6] have been developed.

The HMM-based visual speech synthesis systems that have been developed can be broadly categorized into two types: Image-based systems on the one hand use features derived directly from the video frames [7], [8] where the resulting synthesis is supposed to look like a video of a real person. Motion capture based approaches [9]–[13] on the other hand derive their features from individual facial feature points tracked over time. The advantage of these types of features is that the synthesized motion trajectories can be used to drive any 3D face. Our system is based on motion-capture data but the HSMM-based approach is flexible enough to allow for the synthesis of any type of parameter sequence. Note that our

Manuscript received March 31, 2013; revised July 10, 2013; accepted August 19, 2013. Date of publication September 06, 2013; date of current version March 11, 2014. This work was supported by the by the Austrian Science Fund (FWF): P22890-N23. The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH is funded within the program COMET—Competence Centers for Excellent Technologies by BMVIT, BMWA, and the City of Vienna. The COMET program is managed by the FFG. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Simon King.

D. Schabus is with the Telecommunications Research Center Vienna (FTW), 1220 Vienna, Austria, and also with the Signal Processing and Speech Communication Laboratory, Graz University of Technology, A-8010 Graz, Austria (e-mail: schabus@ftw.at).

M. Pucher and G. Hofer are with the Telecommunications Research Center Vienna (FTW), 1220 Vienna, Austria (e-mail: pucher@ftw.at; hofer@ftw.at).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

This paper has supplemental downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors. The supplementary material consists of an MP4 file showing example stimuli of all three evaluation parts (audio, audiovisual and "challenging" audiovisual) described in the paper. This material is 6.3 MB in size.

Digital Object Identifier 10.1109/JSTSP.2013.2281036

goal is to synthesize both audio speech and motion parameters directly from text, but the models we train can also be used in the less general manner of [7] and [14] (audio unit alignment on speech input followed by visual synthesis, using audiovisual models).

Combining the auditory and visual modalities in one framework requires a synchronous corpus of parametrized facial motion data and acoustic speech data. We have demonstrated in previous work how to build such a corpus [15] and that it is feasible to produce both acoustic speech parameters and animation parameters [16] by a maximum likelihood parameter generation algorithm [17] from models that were trained on such a synchronous corpus. In [18] we showed how to generate visual parameters using a speaker-adaptive approach. The work described in this paper will describe a joint audiovisual speaker-dependent HSMM-based approach for generating visual and acoustic features for different speakers. In statistical data-driven audiovisual synthesis, commonly separate acoustic and visual models are trained [7]–[11], [14], sometimes together with an additional explicit time difference model to correctly synchronize the two modalities [12], [13]. In contrast, we propose to train one joint audiovisual model (with acoustic and visual streams), such that the likelihood of the model generating the training data is maximized globally, across the two modalities, during model parameter estimation. This results in a single duration model used for both modalities, thus eliminating the need for additional synchronization measures. In this way, we intend to create simple and direct models for audiovisual speech synthesis, which can cope with most effects of co-articulation and inter-modal asynchrony naturally through five-state quin-phone full-context modeling. [13] also argues that states can capture some inter-modal asynchrony since transient and stable parts of the trajectories of different modalities need not necessarily be modeled by the same state, and that multi-phone context models can capture co-articulation effects. Notably, an early work on audiovisual HMM-synthesis [19] also applied joint modeling in our sense, however without investigating its benefits in detail. Also, the current HMM-modeling techniques and high-fidelity visual parameter acquisition we use distinguish our work from [19].

Therefore the main purpose of this paper is to investigate whether the proposed joint audiovisual modeling approach provides clear improvements over separate audio and visual modeling. We argue that the main weakness of separate modeling stems from the difficulty to capture (and even define) clear temporal unit borders for the visual modality. Our analysis shows that visual-only training yields models which fail to find suitable borders for some phones when we carry out forced alignment on our training data. An explicit audio/video lag model used for modality synchronization, which is trained on such borders (as in [12], [13]) might still suffer from these problems, even if the borders in the training data are hand-labeled (as in [20]). Furthermore, the quality of the synthesized trajectories themselves can be expected to degrade if observation assignment to units is unclear during training.

On the other hand, there are situations where the targets to which speech needs to be synchronized are much clearer, like singing synthesis [21], where explicit lag models have been used successfully for synchronizing speech to sheet music (in that case, the sheet music defines fixed and exact synchronization target points in time).

We furthermore consider the description of the system we have built an important part of this work. This system is based on a state-of-the-art HSMM modeling framework and we use current animation-industry-standard motion tracking and character animation technology for the visual modality. In this regard our work differs strongly from conceptually related previous work [8]–[10].

The remainder of this paper is organized as follows: Section II describes our data and synthesis systems. Section III provides an analysis of acoustic and visual alignments using the different models. In Section IV we evaluate our different models in subjective listening experiments. Section V concludes the paper.

II. System Description

In this section, we describe the full pipeline of the system we propose, including audiovisual data recording (Section II-A), feature extraction (Section II-B), HSMM training (Section II-C), synchronization strategies (Section II-D) and creation of the final animation (Section II-E).

A. Data

Similar to a corpus we have described before [15], we have recorded four speakers reading the same recording script in standard Austrian German. This script is phonetically balanced, i.e., it contains all phonemes in relation to their appearance in German, and it contains utterances of varying length, to cover different prosodic features (like phrase breaks, etc.). It amounts to 223 utterances and roughly 11 minutes total for each of the speakers.

The recordings were performed in an anechoic, acoustically isolated room with artificial light only. For the sound recordings, we used a high-definition recorder (an Edirol R-4 Pro) at 44.1 kHz sampling rate, 16 bit encoding, and a professional microphone (an AKG C-414 B-TL). We believe this to be sufficient but necessary quality settings, as it has been shown that sampling rates higher than the common 16 kHz can improve speaker similarity in HSMM-based speech synthesis [22].

For the recording of facial motion, we used a commercially available system called OptiTrack [23]. Using six infrared cameras with infrared LEDs, this system records the 3D position of 37 reflective markers glued to a person's face at 100 Hz. A headband with four additional markers helps to segregate global head motion from facial deformation. A seventh camera records 640×480 grayscale video footage, also at 100 Hz (synchronized). See Fig. 1 for still images from the grayscale video showing the marker layout (top), and renderings of the resulting 3D data (bottom). Each recording session was started with a neutral pose (relaxed face, mouth closed, eyes open, looking straight ahead). Using this kind of data (recorded or synthesized), the movement of a virtual 3D head can be controlled as described in Section II-E.

Since our final goal is lip motion synthesis, we have to remove global head motion from the data. This can be done under the assumption of fixed distances between the four headband markers.

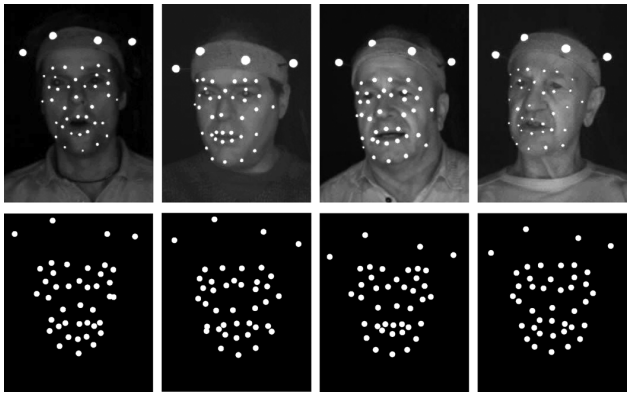


Fig. 1. Still images from grayscale videos showing facial marker layout (top) for four different speakers and corresponding renderings of 3D marker data (bottom).

We choose a reference frame, and compute the transformation matrix from all the other frames to the reference frame, such that the four headband markers are in the same position. By application of this transformation matrix to all 41 markers in the respective frame, we can eliminate global head motion, keeping only the facial deformation in the data.

Furthermore, we have applied a global translation to each recording session's data, such that the head is located at the same position in coordinate space.

B. Visual Feature Extraction

By stacking the x, y and z coordinates of all 41 markers vertically, we obtain 123-dimensional column vectors representing the shape of the face at a given point in time. Because we are interested in the synthesis of speech articulation motion only, we have removed the four headband markers, the four markers on the upper and lower eyelids and the six markers on the eyebrows from the data, resulting in 81-dimensional feature vectors.

Since there are many strong constraints on the deformation of a person's face while speaking, and hence on the motion of the facial markers, there should be far fewer degrees of freedom necessary than these 81-dimensional vectors allow. Guided by this intuition, as well as to de-correlate the components, we have carried out standard principal component analysis (PCA) on our data. We are interested in de-correlation because it will allow us to assume independence between the components and thus to train diagonal rather than full covariance matrices. The other reason for using PCA is that the resulting components are sorted according to their influence on variability in the data, and hence we can choose to keep only the first k principal components instead of the entire 81 dimensions, leading to faster training and more accurate modeling, but still achieve satisfactory results. Appendix A provides a detailed description of how we carried out PCA on our data.

In a recent study [24], we showed that deciding on a value for k based on objective measures such as the singular values or the reconstruction error (see also Fig. 10 in Appendix A) is not straightforward. It is clear that the first dimensions will always explain most of the variance in the data (by the nature of PCA), but deciding on a value for k that will still include even subtle speech motion might require thorough subjective evaluations. A subjective experiment in [24] with a speaker-adaptive setting in mind showed that up to 30 dimensions can be necessary for robust reconstruction. The optimal value for k for training and synthesis may be lower than that, and the system of Bailly *et al.* [12], [13] for example only uses six degrees of freedom, based on a thorough investigation using facial markers and MRI [25]. However, unlike our setup, that study considers symmetrical facial motion only (as does the audiovisual speech synthesis system of [12] and [13]), and the choice of degrees of freedom was based on objective measures alone. For lack of a tighter bound, we have therefore chosen $k = 30$ for the remainder of this paper.

C. Audio, Visual and Audiovisual Model Training

For training regular audio speech models, we use the CSTR/EMIME TTS system training scripts [26] and HTS version 2.1 [27] to train context-dependent, five-state, multi-stream, left-to-right, Multi-Space Distribution (MSD) Hidden Semi-Markov Models (HSMMs) [28]. As audio features we use 39+1 mel-cepstral features, log F0 and 25 band-limited aperiodicity measures, extracted from 44.1 kHz speech, as it is done in the CSTR/EMIME system. Speech signals are re-synthesized from these features using the STRAIGHT vocoder [29]. All features are augmented by their dynamic features (Δ and Δ^2) [30]. For each of the three audio features, the models are clustered separately state-wise by means of decision-tree based context clustering using linguistically motivated questions on the phonetic, segmental, syllable, word and utterance levels. State durations are modeled explicitly rather than via state transition probabilities (HSMMs rather than HMMs [31]), and duration models are also clustered using a single decision-tree across all five states. The feature questions used for the clustering are based on the English question set in the EMIME system [26] with adaptations towards our German phone set. They are listed in [32], except that we do not use multiple dialects here and that we also included the PEC/viseme classes of preceding, current, and succeeding phones (as described below).

In short, for *audio-only* modeling, we apply the state-of-the-art CSTR/EMIME HTS system without modifications. For *visual-only* modeling, we use the same system but with only one feature stream for the visual PCA-space features described in the previous subsection. In order to obtain the same frame rate as the audio features (5 ms frame shift, i.e., 200 frames per second), we have up-sampled (interpolated) the visual features from their native 100 frames to 200 frames per second. Similar to the cepstral features, they are also augmented by their dynamic features and the models are clustered using the same set of questions. This results in a speaker-dependent text-to-visual speech system, like we have investigated in previous work [18]. Furthermore, for *joint audiovisual* modeling, we merge the two into a system that trains models for the three audio features (cepstral, F0, aperiodicity) and the visual features simultaneously. This is achieved by adding an additional stream to the audio-only system, with separate state-wise clustering. The structure of the audio, visual and audiovisual systems is shown in Fig. 2.

As we have added an additional non-standard feature to the well-established HSMM training system, it is of interest to see

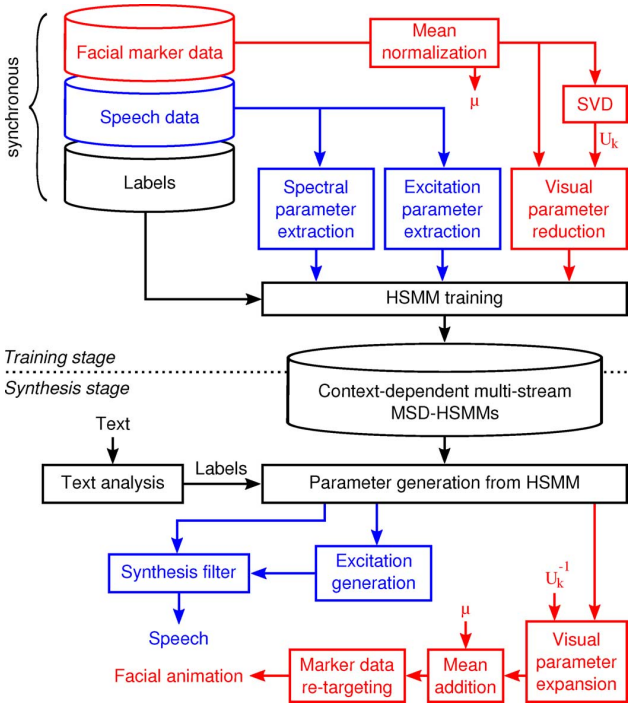


Fig. 2. Overview of a speaker dependent audiovisual speech synthesis system, which consists of three main components: audiovisual speech analysis, audiovisual training, and audiovisual speech generation. The corresponding audio-only system does not include the red parts, and the corresponding visual-only system does not include the blue parts.

TABLE I
AVERAGE NUMBER (ACROSS FOUR SPEAKERS) OF LEAF NODES IN THE CLUSTERING TREES AFTER TRAINING

		State					
Training	Feature	1	2	3	4	5	Total
Audio	Mel-cepstral	58	61	69	66	67	320
	Log F0	146	219	241	149	100	856
	Band-Ap	27	34	36	30	25	152
	Duration						163
Audiovisual	Mel-cepstral	57	63	67	58	61	306
	Log F0	164	218	259	164	121	925
	Band-Ap	27	31	32	23	27	140
	Visual	258	526	551	417	291	2042
	Duration						208
Visual	Visual	354	504	418	345	314	1934
	Duration						312

how the new feature is handled by the system. One potentially informative parameter for this is the size of the clustering trees. Table I gives the number of leaf nodes (and hence of distinct observation probability density functions) resulting from the audio, audiovisual and visual training procedures, averaged across the four speakers. The absolute numbers in such a table of course grow with the size of the training corpus, but we can observe that the trees for the visual features are substantially bigger than the ones for the other features, which is still true if we choose a different dimensionality k to represent our visual data, as illustrated in Fig. 3 where the number of visual leaf

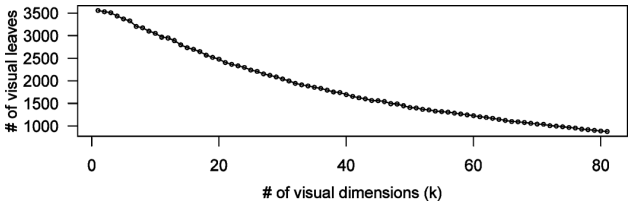


Fig. 3. Average number (across four speakers) of total leaf nodes in the visual clustering trees as a function of visual PCA dimensions kept (k).

nodes is shown as a function of k resulting from audiovisual training. This is somewhat surprising, given that the visual parameter trajectories appear to be quite smooth in general (see Fig. 5 for an example). We interpret this as a strong dependency on context of our visual data.

We also find that the size of the duration tree of the visual-only voice model is roughly twice the size of the audio-only duration tree, and that in the combined audiovisual system we also see an (albeit smaller) increase in size of the duration tree. Duration and audiovisual synchronization will be discussed in more detail in Sections II-D and III, but we can already see from these numbers that duration modeling for the visual features seems to work differently from the audio features.

In many approaches to (audio-)visual speech processing, the concept of *visemes* [33]–[35] or, more generally, Phoneme Equivalence Classes (PECs) [36] is used. The idea is roughly that phone(me)s which have similar or even indistinguishable visual appearance (but which may still be very different in acoustic terms) are grouped together for visual modeling. It is easy to integrate this concept into the HSMM modeling framework, even with the flexibility to use the concept only partially: By "offering" to the model clustering algorithm additional questions that correspond to such groupings of phones according to their visual properties, the maximum description length criterion will automatically make use of such PEC questions when and only when they are useful. To determine to what degree PECs are beneficial or even necessary for visual speech modeling in our setting, it is therefore sufficient to simply provide additional questions alongside the ones mentioned earlier (e.g., phones and phone groups based on acoustic criteria) and then to see whether these are used to cluster the data at hand.

Based on the "easy set" in [36], with adaptations towards our phone set for German, we have added the following six PECs as possible clustering questions: {p, b, m}, {f, v}, {t, d, s, z}, {k, g, n, ŋ, l, h, j, ç, x}, {oː, uː, yː, øː}, {ɔ, ʊ, ʏ, œ}.

Assuming that such PECs are useful for modeling the visual features but not the acoustic ones, these questions should appear often in the clustering trees for the former and rarely (or not at all) for the latter, when they are "offered" at all clustering steps of all features. The percentages of decision tree leaves affected by PEC questions are given in Table II for the three training procedures and all features, averaged across four speakers. Here we consider a leaf "affected" if at least one PEC question was answered affirmatively on the path from the root to the leaf. In line with the expectations mentioned before, we see that PEC questions clearly play a more important role in clustering the models for the visual features than for the acoustic ones, although they

TABLE II
AVERAGE PERCENTAGE (ACROSS FOUR SPEAKERS) OF LEAF NODES AFFECTED BY PEC QUESTIONS

		State					
Model	Feature	1	2	3	4	5	Overall
Audio	Mel-cepstral	8.9	6.1	5.8	5.4	7.6	6.7
	Log F0	7.9	5.9	4.9	3.3	4.9	5.4
	Band-Ap	1.0	4.2	3.9	2.6	0.0	2.5
	Duration						5.1
Audiovisual	Mel-cepstral	9.3	4.8	4.6	1.7	5.4	5.1
	Log F0	8.1	7.4	7.0	6.0	6.2	7.0
	Band-Ap	6.1	3.3	0.7	2.2	2.7	2.9
	Visual	13.1	10.2	22.3	26.2	13.5	17.3
	Duration						12.7
Visual	Visual	13.9	26.4	22.9	26.7	17.5	21.9
	Duration						13.1

are also used for the latter to some extent. PEC questions are especially relevant for the third (22.3%) and fourth (26.2%) states of the visual stream. Interestingly, the presence of the visual features also has an impact on the duration clustering in this respect (in addition to making the duration trees larger, as we have discussed earlier): The duration trees of the visual-only and the audiovisual models contain a higher percentage of PEC-affected leaves than the acoustic-only models.

We conclude from these findings that the addition of clustering questions specifically targeted towards visual features such as visemes or PECs can be helpful in modeling the visual modality in this framework.

D. Audiovisual Synchronization Strategies

To achieve the goal of text-to-audiovisual-speech synthesis, both an acoustic speech signal and a visual speech signal (animation) need to be created given some input text, and in addition to being natural or believable individually, the two generated sequences need to *match temporally*. With the three trained models described in the previous subsection available (audio-only, visual-only and joint-audiovisual, each with its own duration model), there are several possible strategies that lead to a combined audiovisual sequence generated for some new input text.

1) Unsynchronized: The simplest strategy using the separately trained models is to synthesize from each model independently and then just add the two generated sequences together. This has the advantage that each model will generate its sequence "naturally," i.e., the way that directly emerges from the training process of the respective model. An important disadvantage is that there are no synchronization constraints whatsoever, and the total length of the generated audio and visual sequences may even differ. We will refer to this method, which uses two duration models, as *unsync* for short.

2) Utterance Length (Audio): While still using both duration models, we can ensure equal sequence length by adjusting the speaking rate parameter ρ in the synthesis step [37]. The state durations of an utterance consisting of K states (i.e., $K/5$ phones) are given by

$$d_A(k) = \mu_A(k) + \rho \cdot \sigma_A^2(k) \qquad \text{for } 1 \le k \le K, \tag{1}$$

where $\mu_A(k)$ and $\sigma_A^2(k)$ denote the mean and variance of the audio duration model for state k, respectively. When ρ is set to 0 for synthesis, we obtain speech in average speaking rate, with $\rho < 0$ we obtain faster and with $\rho > 0$ slower speech. We can synthesize acoustically without constraints ($\rho_A = 0$), and then determine the ρ_V required for visual synthesis that will yield the same utterance length:

$$D_A = \sum_{k=1}^{K} d_A(k) = \sum_{k=1}^{K} \mu_A(k) \tag{2}$$

$$\rho_V = \frac{D_A - \sum_{k=1}^{K} \mu_V(k)}{\sum_{k=1}^{K} \sigma_V^2(k)}, \tag{3}$$

where $\mu_V(k)$ and $\sigma_V^2(k)$ denote the mean and variance of the visual duration model for state K.

This will produce an audio and visual parameter sequence for the utterance which are exactly of the same length, but still each use their respective duration model. We will refer to this strategy, which exhibits the "natural" audio duration, as *uttlen-audio* for short.

3) Utterance Length (Visual): Symmetrically, by flipping the roles of audio and visual models, we obtain another strategy that exhibits the "natural" visual duration, referred to as *uttlen-visual*.

4) Audio Duration Copy: In order to achieve tighter synchronization on the phone level, we can decide to use only one of the two duration models, e.g., the audio duration model for both audio and visual synthesis. This is equivalent to replacing the visual duration models and trees with the ones obtained from audio training. The advantage here is the tighter synchronization, a possible disadvantage is that a new duration model is forced upon the visual system which might not match the visual feature models. We will refer to this strategy as *durcopy-audio*.

5) Visual Duration Copy: Likewise, we can replace the audio duration model with the visual one, which we will call *durcopy-visual*.

6) Joint Audiovisual: Finally, the audiovisual voice model with jointly trained features and with a single audiovisual duration model generates synchronized parameter trajectories implicitly. A priori it is not clear what kind of effect the additional visual stream will have on the quality of the generated audio samples. One can imagine that the additional information will lead to more robust parameter estimation and thus to an improvement of audio quality. On the other hand, if the two signals reveal themselves to be rather inconsistent, a negative effect on audio quality could arise. We will refer to this strategy as *audiovisual*.

The six synchronization strategies are summarized in Table III. Note that the first three (*unsync*, *uttlen-audio*, *uttlen-visual*) use two duration models whereas the last three (*durcopy-audio*, *durcopy-visual*, *audiovisual*) each use a different single duration model. Furthermore note that *unsync*, *uttlen-audio* and *durcopy-audio* produce synthetic speech identical to what the regular audio-only system would produce.

TABLE III
SYNCHRONIZATION STRATEGIES FOR AUDIOVISUAL SYNTHESIS

Name	Description
unsync	unsynchronized separate duration models
uttlen-audio	utterance length determined by audio duration model
uttlen-visual	utterance length determined by visual duration model
durcopy-audio	audio duration model used for both modalities
durcopy-visual	visual duration model used for both modalities
audiovisual	features trained jointly, audiovisual duration model

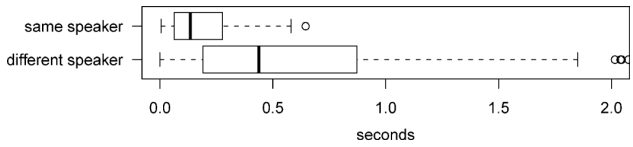


Fig. 4. Boxplots for the differences in utterance length between audio-only and visual-only synthesized utterances. For 23 test utterances and 4 speakers, the top boxplot contains all 92 combinations where audio and visual models were from the same speaker, and the bottom boxplot contains all 276 combinations where the sequences were synthesized using two different speakers' models.

The *unsync* method does not guarantee that audio and visual sequences have the same length, but since both models are trained on the same synchronous corpus, the deviation can be expected to be small, as illustrated in Fig. 4, which shows boxplots of the difference in length when the same utterance is synthesized from an audio-only and from a visual-only model separately. The figure also shows clearly that this difference is significantly smaller when the two models are from the same speaker (and thus trained on a synchronous corpus), suggesting that this synchronization strategy can not work for mixed-speaker setups, if at all.

The *durcopy-audio* method is a straightforward choice to align the borders of both sequences by simply using the borders predicted by the audio model also for the visual model, applied for example in [8] and [16].

The *uttlen-audio* method is interestingly similar to the explicit lag models of [12], [13]: with *uttlen-audio*, audio is synthesized independently of the visual features, and a separate visual duration model predicts the visual phone borders, while the length constraint ensures equal total length of the two sequences. The separate visual model results from several iterations of embedded training on visual-only data. The main difference is that [12], [13] predict the visual phone borders as a relative offset to the audio borders, where the offsets are iteratively re-estimated based on visual forced alignment.

E. Creating the Final Animation

Our synthesis models generate a sequence of motion tracking data. The problem of how to animate a talking head automatically from a sequence of parameters is called retargeting. In this work we used a talking head that is included in our motion tracking system that employs a pre-defined retargeting procedure specific to the facial model. In [38], [39] it was shown how the more general problem of transforming the motions of one talking head onto another talking head can be performed through facial motion cloning. But high-quality retargeting still remains a hard problem for large facial meshes. To exploit the full potential of audiovisual speech synthesis it would be necessary to have a retargeting method that is able to deform any talking head appropriately from the synthesized motion tracking data. We are able to synthesize high-quality motion tracking data trajectories but the visual quality of the final animation also depends on the quality of the retargeting procedure as well as the visual appearance of the head.

III. ALIGNMENT ANALYSIS

This section analyzes the temporal alignment behavior of the different models described in the previous section. Although speech movements and the resulting sounds are synchronous in general, it is not clear a priori whether the borders between phones in the visual speech signal should be the same as in the audio speech signal. For example, at the beginning of an utterance, anticipatory gestures can begin in the speech movement signal well before any audible sound is produced. Although somewhat unnatural, it is commonplace in audio speech synthesis (as well as speech recognition) to define sharp borders between the phones of an utterance and to compensate for co-articulation effects by employing context-dependent modeling strategies (as it is also done in the HTS system we use). Given an acoustic model, such phone borders can be found automatically by forced alignment of the known phone sequence to some speech data.

We have applied HSMM-based forced alignment via the *HSMMAlign* tool from HTS version 2.2 [27] to our training data using the different models we have trained, in order to understand the temporal differences between auditory, visual and joint audiovisual modeling. Given the auditory model and the auditory data, this produces for each of the 200 utterances in the training corpus the most likely phone borders that would make the auditory model generate the speech parameters of this utterance. Likewise for the visual model and data, as well as the audiovisual model and data.

Fig. 5 shows an example sentence with the corresponding forced alignment results. In the first row, the visual-only model was used to align the visual data, the resulting phone borders are designated by black vertical lines. For easier interpretation, the plot shows the Euclidean distances between the central upper lip and central lower lip markers as well as between the left and right mouth corner makers, instead of PCA components. In the third row, the auditory-only model was used to align the auditory data. Here, the first three cepstral features are drawn in red in decreasing thickness and F0 is drawn in green. The low flat portions of the F0 signal represent unvoiced parts (undefined F0). All features have been re-scaled to fit into the same vertical range. The second row combines all features, and the alignment was determined using the joint audiovisual model. The bottom row shows the spectrogram of the utterance. It is apparent that there is a difference between the three resulting alignments.

In order to quantify this temporal alignment difference between the three models, we have computed the alignments for all 200 utterances for all four speakers. Then, to assess the degree of agreement between any two models, we have computed the time percentage of each utterance where the two alignments agree. For an utterance consisting of the phone sequence $(p_1, p_2, \ldots, p_n)$, we compute the agreement percentage

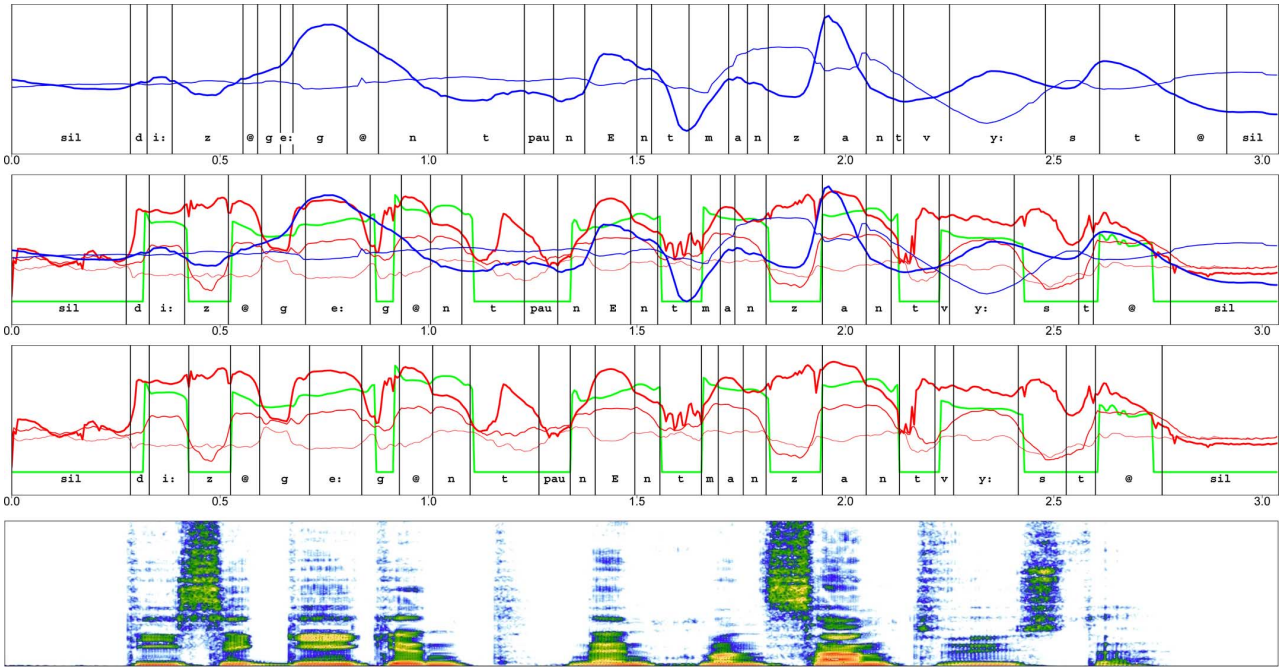


Fig. 5. Result of forced alignment using visual (first row), audiovisual (second row) and audio (third row) models and data. The plots show the Euclidean distance between the central upper lip and central lower lip markers (thick blue line), the Euclidean distance between the left and right mouth corner markers (thin blue line), the first three cepstral features (red, with decreasing thickness) and F0 (green). The different features have been re-scaled to fit into the same vertical range. The bottom row shows the corresponding spectrogram.

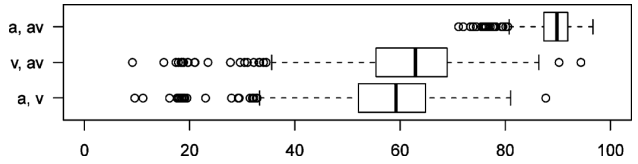


Fig. 6. Boxplots for matching percentage per utterance for audio and audiovisual models (top), visual and audiovisual models (middle) and audio and visual models (bottom).

between two models $A, B \in \{audio, visual, audiovisual\}$ for that utterance as

$$\frac{100}{e_{p_n,A}} \cdot \sum_{i=1}^{n} \max(0, \min(e_{p_i,A}, e_{p_i,B}) - \max(b_{p_i,A}, b_{p_i,B})) \tag{4}$$

where $b_{p_i,X}$ and $e_{p_i,X}$ denote the beginning and the end of phone p_i as determined by *HSMMAlign* using model X. Note that $e_{p_n,A} = e_{p_n,B}$ is simply the total length of the utterance.

The resulting matching percentages of all 800 utterances are shown as boxplots in Fig. 6. The degree of agreement between the auditory and the audiovisual models is much higher (median 89.84%) than between the visual and audiovisual models (median 62.93%) and between the auditory and visual models (median 59.21%). The utterance in Fig. 5 is a typical example in this regard (a-av-match 89.31%, v-av-match 62.66%, a-v-match 58.88%).

We have also computed the matching percentages for any two methods for each individual phone. The percentage is calculated as the amount of time that both alignments consider as being part of the phone divided by the average of the two phone lengths, formally

$$\frac{\max(0, \min(e_{p_i,A}, e_{p_i,B}) - \max(b_{p_i,A}, b_{p_i,B}))}{\frac{1}{2}((e_{p_i,A} - b_{p_i,A}) + (e_{p_i,B} - b_{p_i,B}))} \tag{5}$$

Fig. 7 shows the results grouped by phones (i.e., central phones of the respective quin-phone full-contexts). Apart from the overall better match between auditory and audiovisual (Fig. 7(a)) compared to the two other pairs (Fig. 7(b) and (c)), which is also shown by Fig. 6, it can be seen in these plots that the bottom 12 phones in Fig. 7(b) and (c) are the same, and in almost the same order (by median). These 12 phones show a particularly large mismatch between the visual alignment and both the auditory and the audiovisual alignment, which suggests that for these phones [ə, ʔ, n, t, ɪ, d, g, l, ʀ, ç, h, iː] the training procedure in the visual-only case determined strongly different phone borders from the other two cases. A possible explanation for this is that these phones do not produce prominent effects in the visual feature trajectories, which seems intuitive: since our visual features consist of tracked markers on the lips and face only (and not, e.g., motion features of the tongue or other intra-oral articulators), phones that do not have a strong effect on the movement of the lips and jaw are difficult to capture in the visual feature space. The consonants [ʔ, n, t, d, g, l, ʀ, ç, h] are all mainly defined by intra-oral articulation—in contrast to, e.g., the consonants [f, p, b, m, ʃ], which have a strong effect on lip motion and accordingly appear close to the top in Fig. 7(b) and (c). Likewise, it can be argued that the vowels [ə, ɪ, iː] exhibit rather indistinct lip motion, whereas diphthongs

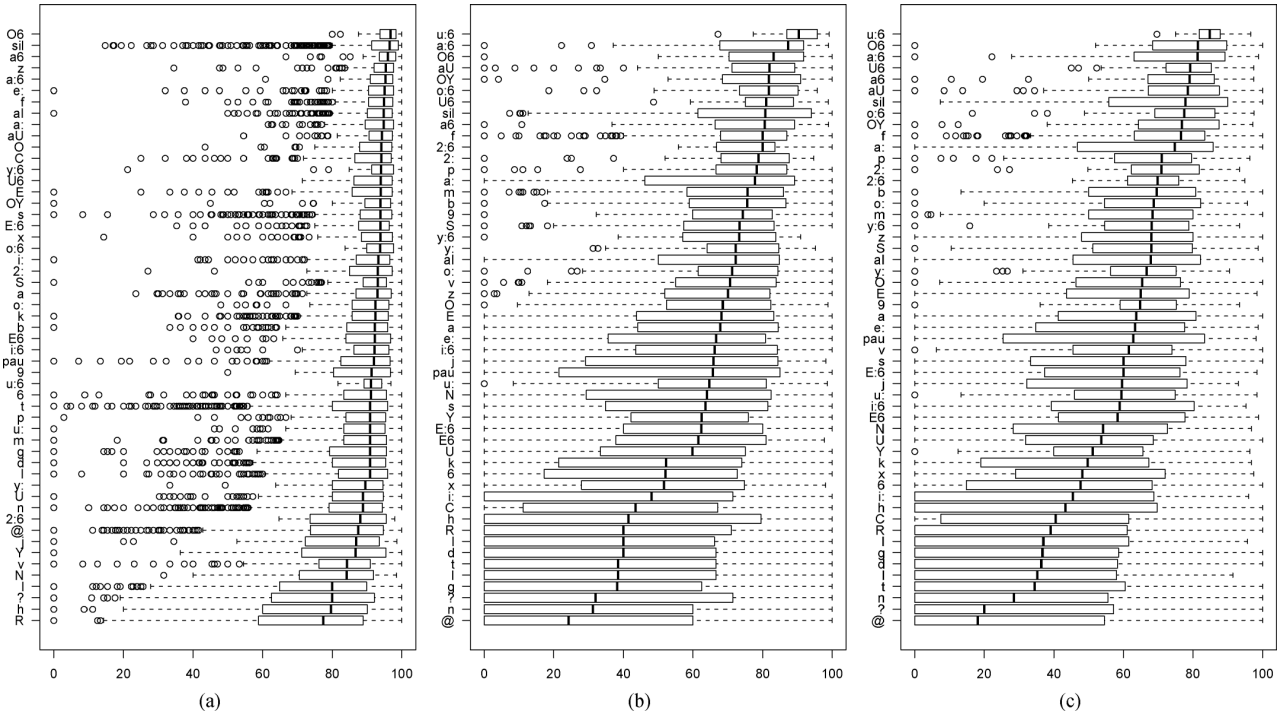


Fig. 7. Boxplots for matching percentage per phone. (a) Auditory and audiovisual. (b) Visual and audiovisual. (c) Auditory and visual.

and rounded vowels can be expected to yield more characteristic trajectories.

IV. Evaluation

In order to assess the quality of the various models and synchronization strategies described in Section II, we have carried out a subjective evaluation experiment with 21 non-expert subjects (13 female, 15 male, aged 20 to 37, mean age 26.5) using a web-based experimental setup. For this experiment, 10 held-out test utterances from our recordings were synthesized using all methods and synchronization strategies and all of our four speakers. The evaluation consisted of an acoustic-only and an audiovisual part.[1]

A. Acoustic Evaluation

To investigate the effect on quality of the audio synthesis of the joint-audiovisual system by adding an additional visual stream, we have evaluated the different methods in a pair-wise comparison listening test. In each comparison, the listeners heard two audio samples from two different methods, but containing the same utterance from the same speaker. After hearing each sample as many times as they liked, they were asked to decide which of the two they preferred with respect to overall quality. No preference (a "tie") was also an option. Four methods for synthesizing audio were compared in this test: *audio*, which represents the regular audio-only system (and hence the synchronization strategies *unsync*, *uttlen-audio* and *durcopy-audio*), *audiovisual*, which represents the audio generated from the joint-audiovisually trained model, *durcopy-visual*, which represents audio synthesized with the visual duration model (used in the synchronization strategy of the same name), and original recorded speech (*recorded*).[2] All possible comparisons were heard twice by different listeners. The results are given in Table IV, where the "winning" scores and the number of ties are listed for each method pair. In the last column, the symbol "$*$" indicates statistical significance of the score difference according to Bonferroni-corrected Pearson's χ^2-tests of independence with $p < 0.01$.

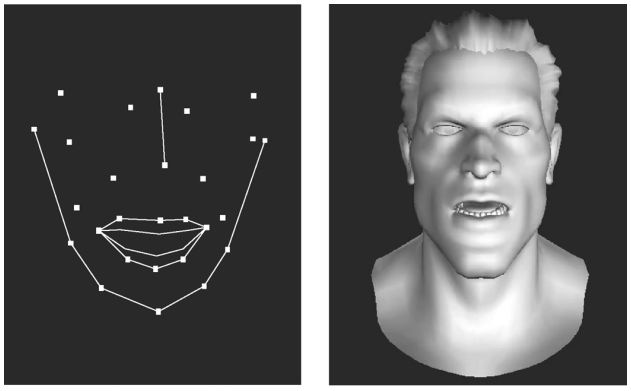
TABLE IV
Evaluation Results for the Acoustic Part

Compared Methods			wins	ties	sig.
recorded	:	audio	76 : 3	1	$*$
recorded	:	audiovisual	77 : 1	2	$*$
recorded	:	durcopy-visual	79 : 1	0	$*$
audio	:	audiovisual	19 : 11	50	
audio	:	durcopy-visual	44 : 6	30	$*$
audiovisual	:	durcopy-visual	43 : 2	35	$*$

Recorded audio was perceived as better than synthetic speech from any of the methods, and audio synthesized using the visual duration model (*durcopy-visual*) was perceived as worse than everything else. The small difference between *audio* and *audiovisual* (19 vs. 11) is not statistically significant ($p > 0.42$) and their similarity is also reflected in the large number of "ties" (50). We interpret these results to indicate that the additional visual stream in the joint audiovisual training has no significant

[1]Example stimuli for all parts of the evaluation are available on http://userver.ftw.at/~schabus/jstsp2013

[2]We did not include the audio from the synchronization strategy *uttlen-visual*, because it is barely if at all distinguishable from *audio*, due to the small absolute values of ρ in our experiments.



(a) (b)

Fig. 8. Mode of speech motion presentation in the first and second (expert) evaluations. Example videos are available on http://userver.ftw.at/~schabus/jstsp2013. (a) Raw marker data. (b) Data-controlled 3D head.

effect (neither positive nor negative) on the quality of the generated acoustic speech signals.

B. Audiovisual Evaluation

In order to evaluate the audiovisual models and in particular the temporal alignment quality of the different synchronization strategies described in Section II-D, we compared rendered videos consisting of synthesized facial motion and synthesized speech in the second part of the experiment. Similar to [40], to focus on evaluating the quality of the generated marker motion rather than the quality of the retargeting procedure or the appearance of the 3D head model, we have decided to present the raw synthesized marker motion to the subjects, i.e., renderings of the 27 points moving in 3D space, with some supporting lines added for orientation as shown in Fig. 8(a). The inner lip contours were added automatically based on a fixed distance between the outer lip markers and six corresponding points that define the inner lip. Even though this method does not necessarily produce all lip closures, it only generates correct lip closures. Note that in a setup with marker motion retargeting to a 3D head, these lines are not needed and all speech motion, including closures and lip compression, is computed based on the marker positions alone by the retargeting procedure.

In each pair-wise comparison in this part of the experiment, the subjects saw two videos from two different methods containing the same utterance from the same speaker. After watching each video as many times as they liked, they were asked to decide which of the two had better synchronization between acoustic speech and visible speech movement. No preference (a "tie") was also an option. We have chosen to ask specifically for synchronization quality, rather than testing more generally for intelligibility and naturalness as it was done in the LIPS 2008/2009 challenges [41].

In this test, we compared all synchronization strategies described in Section II-D, as well as recorded speech and motion data, against each other. The results are given in Table V, where the "winning" scores and the number of "ties" are listed for each method pair. In the last column, the symbol "$*$" indicates statistical significance of the score difference according to Bonferroni-corrected Pearson's χ^2-tests of independence with $p < 0.05$.

TABLE V
EVALUATION RESULTS FOR THE AUDIOVISUAL PART

Compared Methods			wins	ties	sig.
recorded	:	audiovisual	32 : 5	3	$*$
recorded	:	durcopy-audio	25 : 7	8	$*$
recorded	:	durcopy-visual	32 : 6	2	$*$
recorded	:	uttlen-audio	24 : 9	7	$*$
recorded	:	uttlen-visual	26 : 8	6	$*$
recorded	:	unsync	25 : 11	4	$*$
audiovisual	:	durcopy-audio	9 : 17	14	
audiovisual	:	durcopy-visual	18 : 8	14	
audiovisual	:	uttlen-audio	10 : 10	20	
audiovisual	:	uttlen-visual	11 : 20	9	
audiovisual	:	unsync	9 : 14	17	
durcopy-audio	:	durcopy-visual	11 : 9	20	
durcopy-audio	:	uttlen-audio	6 : 11	23	
durcopy-audio	:	uttlen-visual	10 : 12	18	
durcopy-audio	:	unsync	12 : 12	16	
durcopy-visual	:	uttlen-audio	6 : 21	13	$*$
durcopy-visual	:	uttlen-visual	6 : 18	16	$*$
durcopy-visual	:	unsync	8 : 19	13	
uttlen-audio	:	uttlen-visual	8 : 14	18	
uttlen-audio	:	unsync	11 : 9	20	
uttlen-visual	:	unsync	9 : 9	22	

The results in Table V confirm that recorded speech and recorded speech movements were perceived to be synchronized significantly better than any generated stimuli, and that *durcopy-visual* was perceived as having worse synchronization than the two *uttlen* methods. In particular, the *audiovisual* method only performed differently from the *recorded* condition but not from any other method. We expected the *audiovisual* method to be perceived as having the closest synchronization between the visual and the audio stream. However, there are several possible reasons for the absence of such a perceived synchronization:

- The utterances in the evaluation were short (4–7 words), randomly selected held-out test sentences from our recorded data. Longer sentences rich in phones that exhibit prominent lip motion (as identified in Section III) might show stronger differences between the methods.
- The decision to present animated raw marker data rather than an animated 3D head model controlled by this data might have been a counter-productive one.
- The test subjects were non-experts recruited on the web, who might have only reported very obvious differences, resulting in "washed-out" results for the more subtle differences.

To further test the synchronization, an additional evaluation was carried out with subjects judging "challenging" utterances, which were longer (12–17 words), semantically unpredictable but syntactically correct utterances, rich in audiovisual "landmarks," synthesized following the four synchronization strategies *audiovisual*, *uttlen-audio*, *uttlen-visual* and *durcopy-audio*. We do not have recordings of these utterances and we excluded the *durcopy-visual* strategy because of its bad performance in the first evaluation. We also excluded

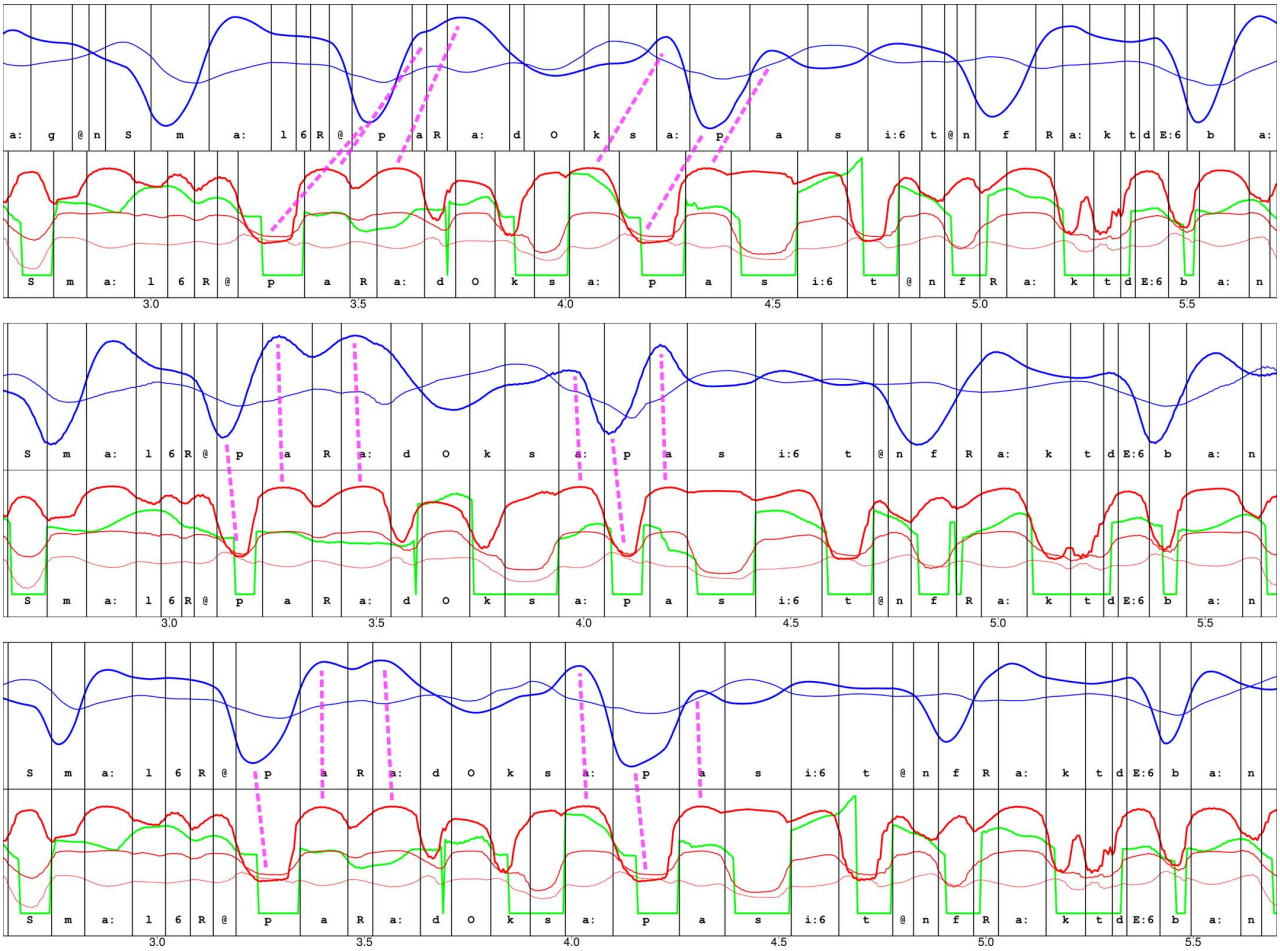


Fig. 9. Excerpts of synthesized audiovisual trajectories for one of the "challenging" utterances from different synthesis strategies: uttlen-visual (top), audiovisual (middle) and durcopy-audio (bottom). The plots show the Euclidean distance between the central upper lip and central lower lip markers (thick blue line), the Euclidean distance between the left and right mouth corner markers (thin blue line), the first three cepstral features (red, with decreasing thickness) and F0 (green). The different features have been re-scaled to fit into the same vertical range. Some feature landmark correspondences are indicated by cyan dashed lines.

TABLE VI
EVALUATION RESULTS USING "CHALLENGING" UTTERANCES

	experts			non-experts		
Compared Methods	wins	ties	sig.	wins	ties	sig.
audiovisual : durcopy-audio	15 : 5	5	*	25 : 24	16	
audiovisual : uttlen-audio	17 : 3	5	*	34 : 22	9	
audiovisual : uttlen-visual	17 : 4	4	*	31 : 23	11	
durcopy-audio : uttlen-audio	10 : 8	7		31 : 15	19	*
durcopy-audio : uttlen-visual	10 : 8	7		30 : 18	17	
uttlen-audio : uttlen-visual	5 : 6	14		18 : 25	22	

unsync because of the strong similarity of this method to the two *uttlen* methods. We applied the synthesized marker motion to a 3D head model via retargeting and created rendered animation sequences from these (see Fig. 8(b) for an example frame). 13 non-expert subjects and 5 expert subjects (speech technology, phonetics) took part in this evaluation (9 female, 9 male, aged 22 to 58, mean age 33.9). Otherwise the experimental setup was identical to the first evaluation. The results are given in Table VI.

For these "challenging" utterances, the experts perceived the *audiovisual* method to produce significantly better speech/motion synchronization than the other methods, which show no significant difference among each other. For the non-expert subjects, on the other hand, the only significant difference is between *ducropy-audio* and *uttlen-audio*. This suggests that the *audiovisual* method produces improved synchronization, but some subtle differences are not consciously perceived by the non-expert subjects, although a clear trend in favor of the *audiovisual* method is also visible for the non-experts.

Fig. 9 shows excerpts of synthesized trajectories for one of the "challenging" utterances. The top part of the figure illustrates the *uttlen-visual* strategy. Although identical total utterance duration is ensured, the two duration models generate different phone durations within the utterance, resulting in a clear misalignment of some feature "landmarks," as indicated in the figure by dashed cyan lines. The middle part of the figure illustrates the joint audiovisual strategy. The single audiovisual duration model provides better alignment of the same feature "landmarks". It is quite obvious that this causes a perceptible improvement over the *uttlen-visual* method. The bottom part il-

lustrates the *durcopy-audio* method. Overwriting the visual duration model with the audio one guarantees alignment of the phone borders, resulting in good alignment also of the feature "landmarks". However, forcing the visual system to use predefined durations can result in artificial contraction or stretching of phones, leading to unnaturally fast or slow movement, as visible in the stretched [p] phone between second 4 and 4.5. As the expert evaluation has shown, this leads to a perceptible inferiority of this synchronization strategy to the *audiovisual* method.

V. Conclusion

In this paper we showed that joint audiovisual speech synthesis improves the quality of the visual speech compared to other synchronization approaches. In our first evaluation we saw no differences between audiovisual modeling and other synchronization approaches, except for the recorded data which was always better than the models. Concerning acoustic synthesis quality, all models except *audiovisual* performed worse than acoustic modeling only.

During an additional evaluation with visually challenging utterances, the audiovisual model performed significantly better than other synchronization approaches when judged by expert listeners. In addition, the analysis of the state-alignments, produced by the different models, showed objective differences in audiovisual alignment between the proposed approaches. In summary the proposed integrated speaker-dependent audiovisual approach allows for joint modeling of visual and acoustic signals while maintaining high-quality acoustic synthesis results with improved audiovisual synchronization over other methods.

A few questions have remained open, mainly because they were not in the main focus of this paper, which we deem interesting for future work. For example, we have seen in Section II-C that the concept of visemes/PECs seems applicable also to joint audiovisual modeling. However, a more extensive investigation including subjective evaluations would be required for a deeper understanding of this topic. Subjective evaluations might also be necessary to decide on an optimal value for the dimensionality of the visual parameters. Furthermore, as we have recorded data from multiple speakers, we would like to investigate mixed-speakers setups and joint audiovisual speaker adaptation. On a broader scale, we see the problem of fully automatic speech motion retargeting as an important remaining challenge for the field of 3D audiovisual speech synthesis, especially concerning lip closures and non-rigid lip deformations. Finally, many applications of audiovisual speech synthesis (e.g., video games, animated films, conversational agents) require believable conversational and emotional synthetic speech, which is still an open challenge for acoustic and even more so for audiovisual speech synthesis.

Appendix
Principal Component Analysis via Singular Value Decomposition

For each speaker, we construct a matrix M of size $81 \times n$ of all frames of all utterances of that speaker stacked horizontally, subtract the sample mean column vector μ from each column of

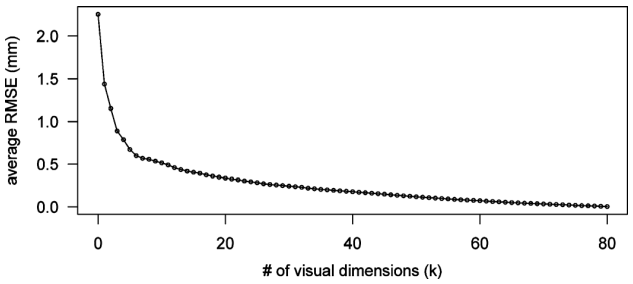


Fig. 10. Average (across four speakers) Root Mean Squared Error (RMSE) of four-fold cross validation PCA reconstruction of visual parameters with varying dimensionality (k).

M to obtain a normalized $\bar{M}$, and compute the Singular Value Decomposition (SVD):

$$\bar{M} = U \cdot \Sigma \cdot V^T \tag{6}$$

We are solely interested in the matrix U of size 81×81, whose columns are the bases of the principal component space, sorted by decreasing eigenvalues. We can project a frame column vector x into principal component space by multiplying U^T from the left ($U^T \cdot x$), and back into the original space by multiplying U^T's inverse from the left. Since U is orthogonal, we have $(U^T)^{-1} = (U^T)^T = U$ and thus

$$x = U \cdot (U^T \cdot x), \tag{7}$$

and if U_k denotes the matrix containing only the first k columns of U, then

$$x \approx U_k \cdot (U_k^T \cdot x), \tag{8}$$

where the quality of the approximation improves with increasing value of k.

So we can carry out SVD on the data M of a speaker, choose a value for $k < 81$ and project the data into a smaller (k-dimensional) subspace using U_k^T. Then, HSMM training and synthesis can be performed using this more compact and de-correlated representation of the speaker's data. Synthesized utterances can be projected back into the full 81-dimensional space using U_k, and by re-adding the sample mean μ we finally obtain the corresponding synthesized facial marker movement.

The influence of k on the quality of the approximation is shown in Fig. 10, which shows the reconstruction error as a function of k from a four-fold cross-validation setup, averaged across four speakers.

References

[1] M. Cohen and D. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, N. Magnenat-Thalmann and D. Thalmann, Eds. New York, NY, USA: Springer-Verlag, 1993, pp. 139–156.

[2] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proc. SIGGRAPH*, Los Angeles, CA, USA, 1997, pp. 353–360.

[3] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *Proc. SIGGRAPH*, San Antonio, TX, USA, 2002, pp. 388–398.

[4] G. Bailly, M. Bérar, F. Elisei, and M. Odisio, "Audiovisual speech synthesis," *Int. J. Speech Technol.*, vol. 6, pp. 331–346, Jan. 2003.

[5] Z. Deng and U. Neumann, "eFASE: Expressive facial animation synthesis and editing with phoneme-isomap controls," in *Proc. Eurographics SCA*, Aire-la-Ville, Switzerland, 2006, pp. 251–260.
[6] B. Theobald, J. Bangham, I. Matthews, and G. Cawley, "Near-videorealistic synthetic talking faces: Implementation and evaluation," *Speech Commun.*, vol. 44, no. 1-4, pp. 127–140, 2004.
[7] L. Wang, Y.-J. Wu, X. Zhuang, and F. K. Soong, "Synthesizing visual speech trajectory with minimum generation error," in *Proc. ICASSP*, May 2011, pp. 4580–4583.
[8] S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "HMM-based text-to-audio-visual speech synthesis," in *Proc. ICSLP*, Beijing, China, Oct. 2000, pp. 25–28.
[9] T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi, and K. Tokuda, "Text-to-visual speech synthesis based on parameter generation from HMM," in *Proc. ICASSP*, May 1998, vol. 6, pp. 3745–3748.
[10] M. Tamura, T. Masuko, T. Kobayashi, and K. Tokuda, "Visual speech synthesis based on parameter generation from HMM: Speech-driven and text-and-speech-driven approaches," in *Proc. AVSP*, Dec. 1998, pp. 221–226.
[11] G. Hofer and K. Richmond, "Comparison of HMM and TMDN methods for lip synchronisation," in *Proc. INTERSPEECH*, Makuhari, Japan, Sep. 2010, pp. 454–457.
[12] O. Govokhina, G. Bailly, and G. Breton, "Learning optimal audiovisual phasing for a HMM-based control model for facial animation," in *Proc. SSW6*, Bonn, Germany, 2007, pp. 1–4.
[13] G. Bailly, O. Govokhina, F. Elisei, and G. Breton, "Lip-synching using speaker-specific articulation, shape and appearance models," *EURASIP J. Audio, Speech, Music Process.*, vol. 2009, no. 769494, pp. 1–11, 2009.
[14] G. Hofer, J. Yamagishi, and H. Shimodaira, "Speech-driven lip motion generation with a trajectory HMM," in *Proc. INTERSPEECH*, Brisbane, Australia, Sep. 2008, pp. 2314–2317.
[15] D. Schabus, M. Pucher, and G. Hofer, "Building a synchronous corpus of acoustic and 3D facial marker data for adaptive audio-visual speech synthesis," in *Proc. LREC*, Istanbul, Turkey, May 2012, pp. 3313–3316.
[16] D. Schabus, M. Pucher, and G. Hofer, "Simultaneous speech and animation synthesis," in *ACM SIGGRAPH '11 Posters*, Vancouver, BC, Canada, 2011.
[17] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, Istanbul, Turkey, Jun. 2000, pp. 1315–1318.
[18] D. Schabus, M. Pucher, and G. Hofer, "Speaker-adaptive visual speech synthesis in the HMM-framework," in *Proc. INTERSPEECH*, Portland, OR, USA, 2012, pp. 979–982.
[19] M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi, "Text-to-audio-visual speech synthesis based on parameter generation from HMM," in *Proc. EUROSPEECH*, Budapest, Hungary, Sep. 1999, pp. 959–962.
[20] L. Terry, "Audio-visual asynchrony modeling and analysis for speech alignment and recognition," Ph.D. dissertation, Northwestern Univ., Chicago, IL, USA, 2011.
[21] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "An HMM-based singing voice synthesis system," in *Proc. Interspeech*, Pittsburgh, PA, USA, 2006, pp. 2274–2277.
[22] J. Yamagishi and S. King, "Simple methods for improving speaker-similarity of HMM-based speech synthesis," in *Proc. ICASSP*, Dallas, TX, USA, 2010, pp. 4610–4613.
[23] Naturalpoint, 2013 [Online]. Available: http://www.naturalpoint.com/optitrack/
[24] D. Schabus, M. Pucher, and G. Hofer, "Objective and subjective feature evaluation for speaker-adaptive visual speech synthesis," in *Proc. AVSP*, Annecy, France, Sep. 2013, pp. 37–42.
[25] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images," *J. Phonetics*, vol. 30, no. 3, pp. 533–553, 2002.
[26] J. Yamagishi and O. Watts, "The CSTR/EMIME HTS system for Blizzard challenge 2010," in *Proc. Blizzard Challenge Workshop*, Kansai Science City, Japan, 2010.
[27] K. Tokuda, H. Zen, J. Yamagishi, A. Black, T. Masuko, S. Sako, T. Toda, T. Nose, and K. Oura, "The HMM-Based Speech Synthesis System (HTS)," 2008 [Online]. Available: http://hts.sp.nitech.ac.jp
[28] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.
[29] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.
[30] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP*, May 1995, pp. 660–663.
[31] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *Proc. ICSLP*, 2004, pp. 1397–1400.
[32] M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, and V. Strom, "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis," *Speech Commun.*, vol. 52, no. 2, pp. 164–179, 2010.
[33] C. G. Fisher, "Confusions among visually perceived consonants," *J. Speech Hear. Res.*, vol. 11, pp. 796–804, 1968.
[34] T. Chen, "Audiovisual speech processing," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 9–21, Feb. 2001.
[35] D. Massaro, M. Cohen, M. Tabai, J. Beskow, and R. Clark, "Animated speech: Research progress and applications," in *Audiovisual Speech Process.*, G. Bailly, P. Perrier, and E. Vatikiotis-Bateson, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2012, pp. 309–345.
[36] L. E. Bernstein, "Visual speech perception," in *Audiovisual Speech Processing*, G. Bailly, P. Perrier, and E. Vatikiotis-Bateson, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2012, pp. 21–39.
[37] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration modeling for HMM-based speech synthesis," in *Proc. ICSLP*, 1998, pp. 29–32.
[38] M. Fratarcangeli, M. Schaerf, and R. Forchheimer, "Facial motion cloning with radial basis functions in MPEG-4 FBA," *Graphical Models*, vol. 69, no. 2, pp. 106–118, 2007.
[39] I. S. Pandzic, "Facial motion cloning," *Graphical Models*, vol. 65, no. 6, pp. 385–404, 2003.
[40] G. Bailly, G. Gibert, and M. Odisio, "Evaluation of movement generation systems using the point-light technique," in *Proc. IEEE Workshop Speech Synth.*, 2002, pp. 27–30.
[41] B.-J. Theobald, S. Fagel, G. Bailly, and F. Elisei, "Lips2008: Visual speech synthesis challenge," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 2310–2313.

Dietmar Schabus received the B.Sc. and M.Sc. degrees in computer science from the Vienna University of Technology, Austria in 2006 and 2009, respectively.

Since 2009, he has been a Researcher at the Telecommunications Research Center Vienna (FTW), Austria. He is currently pursuing the Ph.D. degree in the field of audiovisual speech synthesis at FTW and Graz University of Technology, Austria.

Michael Pucher received a Ph.D. degree in electrical engineering from Graz University of Technology, Austria in 2007.

He is a Senior Researcher and Project Manager at the Telecommunications Research Center Vienna (FTW). He has authored and co-authored more than 40 refereed papers in international conferences and journals. A list of publications and a detailed CV can be found on http://userver.ftw.at/~pucher.

Gregor Hofer obtained his Ph.D. degree in informatics from the University of Edinburgh, United Kingdom in 2009.

He has previously held research positions at the University of Edinburgh and is currently a Senior Researcher at the Telecommunications Research Center Vienna (FTW), Austria. In 2010 he co-founded Speech Graphics to commercialize speech-driven facial animation.

Speaker-adaptive visual speech synthesis in the HMM-framework

Dietmar Schabus[1,2], *Michael Pucher*[1], *Gregor Hofer*[1]

[1]FTW Telecommunications Research Center Vienna, Austria
[2]Graz University of Technology, Graz, Austria
schabus@ftw.at, pucher@ftw.at, hofer@ftw.at

Abstract

In this paper we apply speaker-adaptive and speaker-dependent training of hidden Markov models (HMMs) to visual speech synthesis. In speaker-dependent training we use data from one speaker to train a visual and acoustic HMM. In speaker-adaptive training, first a visual background model (average voice) from multiple speakers is trained. This background model is then adapted to a new target speaker using (a small amount of) data from the target speaker. This concept has been successfully applied to acoustic speech synthesis. This paper demonstrates how model adaption is applied to the visual domain, synthesizing animations of talking faces. A perceptive evaluation is performed, showing that speaker-adaptive modeling outperforms speaker-dependent models for small amounts of training / adaptation data.
Index Terms: Visual speech synthesis, speaker-adaptive training, facial animation

1. Introduction

The goal of audio-visual text-to-speech synthesis is to generate both an acoustic speech signal as well as a matching animation sequence of a talking face, given some unseen text as input. Most commonly, acoustic and visual synthesis are addressed separately, and although we are currently also investigating joint audio-visual modeling, we follow the separated approach in this paper.

Proposed visual speech synthesis systems can be classified according to several criteria, one of them being the distinction between image-based video-realistic methods and model-based 3D methods. While the image-based methods (e.g., [1], [2], [3]) can produce quite convincing results, they often lack flexibility in terms of appearance and perspective, a flexibility that is very desirable in some applications like computer games and 3D-animated films. 3D methods (e.g., [4], [5], [6], [7]), on the other hand, provide this flexibility straightforwardly, but generating convincing speech movements on a 3D face model is challenging. Another possible classification is that of concatenative vs. generative methods, similar to the distinction between the two most common *acoustic* synthesis methods today (unit selection and HMM-based). Our work belongs to the 3D generative group, the details of our pipeline are described in the next section.

However, as with all HMM-based approaches, large amounts of training data are required to build high quality systems and recording large amounts of video data is even more costly than recording audio data. To address this shortcoming for speakers where limited amounts of data are available, a very successful speaker-adaptive approach has been developed [8, 9] for acoustic HMM-based speech synthesis. A (possibly large) speech database containing multiple speakers is used to train an average voice, where a speaker-adaptive training (SAT) algorithm provides speaker normalization. Then, a voice for a new target speaker can be created by transforming the models of the average voice via speaker adaptation, using (a possibly small amount) of speech data from the target speaker. This allows the creation of many speakers synthetic voices without requiring large amounts of speech data from each of them. It can be shown that synthetic speech from voice models created in this way is perceived as more natural sounding than synthetic speech from speaker-dependent voice models using the same (target speaker) data [8]. This holds especially for the case where this data is of small amount. The goal of this paper is to demonstrate how this speaker-adaptive training approach can be applied to visual speech synthesis.

The following Section 2 first describes our data and facial animation pipeline, and then the speaker-adaptive visual speech synthesis system that we have developed, using the acoustic speaker-adaptive system [10] as a basis. We evaluate our system and discuss the results in Section 3. Finally, Section 4 gives a summary and conclusions.

2. Adaptive visual speech synthesis system

2.1. From recorded data to 3D animation

We have recorded a synchronous corpus of acoustic and 3D facial marker data [11], which consists of three speakers of Austrian German, each reading the same 223 phonetically balanced utterances. In addition to high quality audio recordings, we have recorded the 3D positions of 41 reflective markers glued to the speakers' faces at

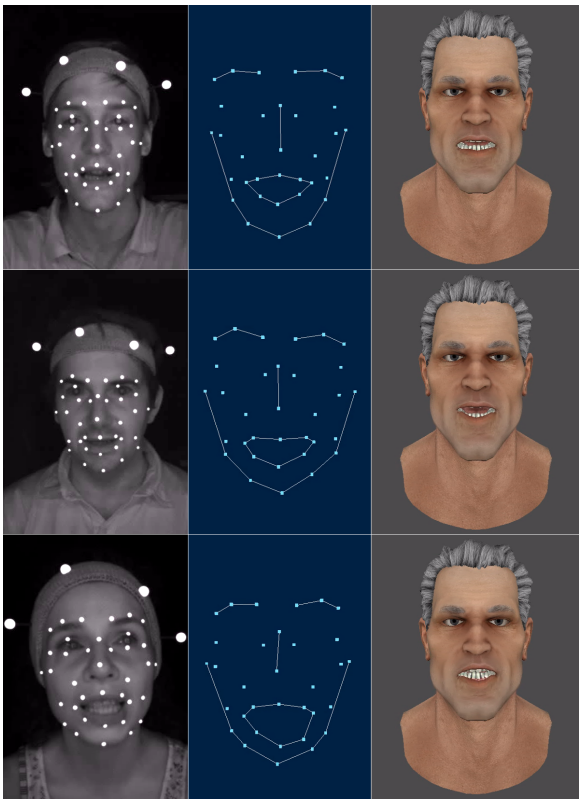


Figure 1: *Still images from the recording session (left), the corresponding 3D marker data (middle) and the resulting pose of the virtual head with this data applied (right). See also videos at http://userver.ftw.at/~schabus/interspeech2012/*

100Hz using a commercially available motion capture system called OptiTrack[1]. This kind of data are commonly used in 3D animation to drive virtual characters within animation software packages. Figure 1 shows still images from the grayscale videos that OptiTrack also records (left), the corresponding 3D marker data (middle) and the resulting pose of the virtual head with this data applied (right).

Global head motion is removed from the data using four headband markers, which become static after "subtracting" head motion and are thus removed from the data. We have also removed the four markers corresponding to the upper and lower eyelids, because we believe phones are inappropriate temporal units for eye blink synthesis. We are thus working with $(41 - 4 - 4) \cdot 3 = 99$-dimensional face representations.

To further reduce dimensionality as well as to achieve de-correlation of the visual features before training, we apply standard principal component analysis (PCA) via singular value decomposition (SVD). HMM-training, adaptation and synthesis are carried out in a k-dimensional PCA subspace of the full 99-dimensional space. After parameter generation in the synthesis step, however, we re-project from the reduced PCA space into the original 99-dimensional space. Therefore the final output of our system has the same format as the data originally recorded. In this way, our method generalizes to different marker layouts, head models and even marker motion re-targeting methods.

We have analyzed the features produced via PCA using objective reconstruction error calculations [11] as well as a perceptive evaluation. Based on those results, we have decided on $k = 30$, i.e., we operate in a 30-dimensional subspace of the full 99-dimensional space.

Given a (recorded or synthesized) sequence of marker positions, we drive a 3D head model with matching control points (called the *rig* or the *bones* in animation terminology) within a professional animation software, and generate rendered video clips from there.

Our corpus also contains HTK quin-phone full-context label files, providing the transcription with precise temporal phone borders. The borders were determined by carrying out hidden-Markov-model based flat-start forced alignment on the acoustic data. We are aware that the temporal borders of phones are not necessarily identical in acoustic and visual data, and that there have been efforts to address exactly these discrepancies [12], but in our experience context-dependent phone modeling seems to already alleviate this problem.

2.2. Visual parameter modeling framework

Figure 2 shows the speaker-adaptive visual modeling framework. The whole system consists of a training, adaptation, and synthesis module. Context-dependent, left-to-right, hidden semi-Markov models (HSMMs) are trained on multi-speaker visual databases in order to simultaneously model the visual features, as well as duration. We use speaker-adaptive training (SAT) based on constrained maximum likelihood linear regression (CMLLR) for the training of the average visual models [8, 9].

The visual feature extraction is applied to a multi-speaker database before training, and to a possibly different single speaker database before adaptation. In the synthesis step, visual parameters are generated from the adapted models.

The visual feature extraction for the training of the average visual voice first applies mean normalization and SVD to derive a matrix U_k that is used to project the data to a lower k-dimensional space. In the adaptation step we also perform mean normalization using the speaker mean μ_s and then use U_k from average voice training to reduce the visual adaptation features. In visual synthesis, the generated features are projected back to the full feature space using U_k^{-1}, and the speaker mean μ_s is added. The resulting visual features are used to animate a talking

[1] http://www.naturalpoint.com/optitrack/

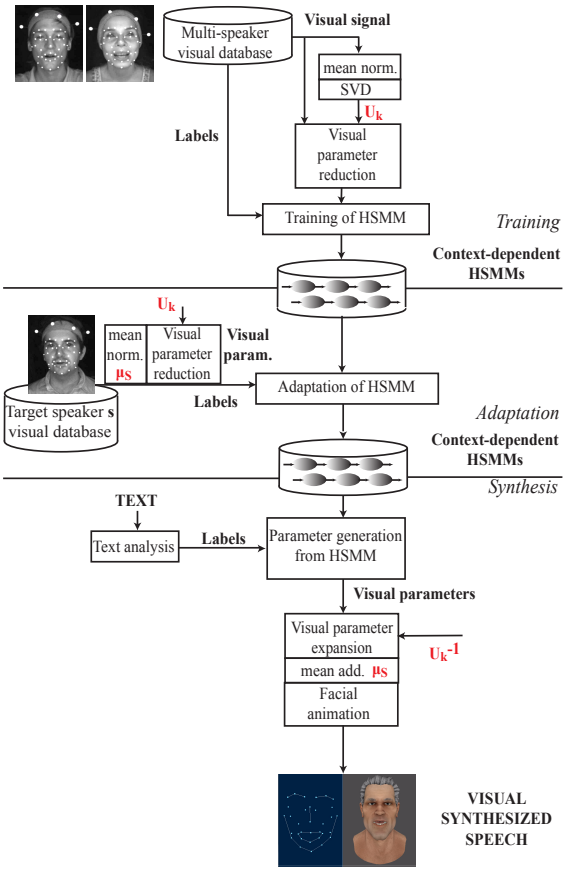


Figure 2: *Overview of the speaker-adaptive visual modeling framework.*

head.

We would like to emphasize that the feature projection matrix U_k is the same in the training, adaptation and synthesis steps, and that it is determined via SVD without using data from the target speaker, i.e., in the entire process there is only one SVD calculation, namely across all speakers that contribute to the average voice. The speaker means, on the other hand, are subtracted per speaker before SVD and projection in the training part, and also before projection in the adaptation part.

In speaker-dependent modeling, the training data comes from one speaker s, U_k and μ_s are determined on that speaker's data and the whole adaptation step is missing.

3. Evaluation

To evaluate our system, 10 held-out test utterances where synthesized. In order to allow for direct comparison of recorded data to synthesized utterances, the true phone durations from the recorded data were employed instead of generated durations from the trained duration models.

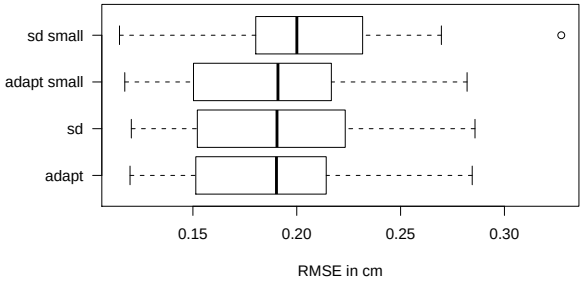


Figure 3: *Box plots of the root mean squared differences between synthesized and recorded marker positions.*

This results in all stimuli from the same speaker and utterance to be of equal length on a phone-by-phone basis.

We compare the recorded data (which we will refer to as *recorded*) to four training strategies: 1. The speaker-adaptive method we presented in the previous section, where an average voice is trained on the data of two speakers (212 utterances each), which is then adapted to the third speaker using also 212 utterances (*adapted*). 2. A corresponding speaker-dependent model, trained on the target speaker's 212 utterances (*sd*). 3. An adapted model with a small amount of adaptation data; here, the average voice is the same as in *adapted*, but for adaptation we use the smallest set of utterances that contains each phone at least three times (19 utterances) (*adapt small*). 4. A speaker-dependent model trained on the same small amount of data (*sd small*).

Similar to our objective reconstruction error calculations during the analysis of the PCA projections, we have computed objective errors by calculating the frame-wise deviations of marker positions between recorded and synthesized sequences. Figure 3 shows the resulting root mean square errors (RMSE), calculated across all frames of each utterance. Since we have 10 test utterances and three speakers, each box plot contains 30 RMSE values. Unfortunately, these objective results are not very informative. If anything, we can observe that the RMSE for *sd small* is slightly larger than for the other methods. This is mainly due to temporal misalignment: although we force the parameter generation to produce the same phone durations as the ones in the recorded data, slight temporal shifts of the valleys and peaks of a trajectory within a phone can cause a large error even though the movement of the corresponding maker is "correct". Objective evaluation of synthesized marker motion by comparison to recorded data is therefore not straightforward.

Therefore, we have conducted a perceptive experiment with 28 test subjects (11 female, 17 male, aged 15 to 49, mean age 27.5). Each subject saw 45 pairs of videos showing a virtual head driven by two different models (*recorded*, *sd*, *sd small*, *adapted*, *adapted small*), where all possible combinations of methods, speakers and utterances were distributed among the subjects such that each

Table 1: *Pair wise comparison scores*

Compared methods			wins	ties	sig.
recorded	:	sd	74 : 33	20	$*$
recorded	:	sd small	95 : 25	10	$*$
recorded	:	adapt	95 : 20	10	$*$
recorded	:	adapt small	86 : 22	10	$*$
sd	:	sd small	64 : 36	22	$*$
sd	:	adapt	54 : 39	28	
sd	:	adapt small	56 : 37	39	
sd small	:	adapt	56 : 34	35	
sd small	:	adapt small	31 : 57	37	$*$
adapt	:	adapt small	27 : 35	73	

subject saw each of the ten method combinations, as well as each speaker-utterance at least once. To each video we have added a synthetic speech sample generated from models that we trained on the corresponding speaker's acoustic data from our synchronous corpus. As for the visual synthesis, we have provided the phone borders from the recordings rather than using the duration model.[2]

For each video pair, the subjects selected whether they preferred the first or the second video, or they thought they were of equal quality. The results are given in Table 1, where we have counted the number of "won" comparisons and the number of "ties" for each method pair. To assess the statistical significance of these preference scores, we have computed Bonferroni-corrected Pearson's χ^2-tests of independence with $p < 0.01$ for each method pair. The results are given in the last column of Table 1, where the symbol "$*$" indicates a statistically significant influence of the methods on the preference scores.

The animations that replay the recorded data are preferred significantly more times over all the synthesis methods. Furthermore, within the speaker-dependent methods *sd* and *sd small* the reduction in training data results in a significant difference between the two. The result between *sd* and *adapt* is not significant, but shows a trend towards the speaker-dependent model. However, *adapt small* is preferred over *sd small*, and the difference is statistically significant.

4. Conclusion

All in all this work demonstrated how to apply average voice training and speaker adaptation to visual speech synthesis. This is useful when creating new systems for speakers where very few training utterances are available. In addition with limited amount of training data the speaker adaptive approach outperforms speaker dependent training. However, several additional experiments will be conducted in future work. In particular speaker similarity, a measure of how close synthesized data mimics specific speaker characteristics, will be investigated. We are also currently recording a large multi-speaker audio-visual database of different dialects of Austrian German. Further work will address how to apply the methods developed in this paper to more speakers and more training data.

[2] See example stimuli at http://userver.ftw.at/~schabus/interspeech2012/

5. Acknowledgements

We would like to thank Junichi Yamagishi for help with HTS and Priska Lang for finding the test subjects.

This research was funded by the Austrian Science Fund (FWF): P22890-N23. The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH is funded within the program COMET – Competence Centers for Excellent Technologies by BMVIT, BMWA, and the City of Vienna. The COMET program is managed by the FFG.

6. References

[1] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *ACM SIGGRAPH*, New York, NY, USA, 2002, pp. 388–398.

[2] J. Melenchon, E. Martinez, F. De La Torre, and J. Montero, "Emphatic visual speech synthesis," *IEEE Trans. Audio, Speech, & Language Proc.*, vol. 17, no. 3, pp. 459–468, 2009.

[3] L. Wang, Y.-J. Wu, X. Zhuang, and F. K. Soong, "Synthesizing visual speech trajectory with minimum generation error," in *Proc. ICASSP*, 2011, pp. 4580–4583.

[4] F. Parke, "A parametric model of human faces," Ph.D. dissertation, University of Utah, Salt Lake City, UT, USA, 1974.

[5] J. Beskow, "Talking heads – models and applications for multimodal speech synthesis," Ph.D. dissertation, KTH Stockholm, 2003.

[6] T. H. Chen and D. W. Massaro, "Evaluation of synthetic and natural mandarin visual speech: Initial consonants, single vowels, and syllables," *Speech Communication*, vol. 53, no. 7, pp. 955–972, 2011.

[7] S. Fagel and G. Bailly, "German text-to-audiovisual-speech by 3-D speaker cloning," in *Proc. AVSP*, Tangalooma, QLD, Australia, Sept 2008.

[8] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[9] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech & Language Proc.*, vol. 17, no. 1, pp. 66–83, 1 2009.

[10] J. Yamagishi, T. Nose, H. Zen, T. Toda, and K. Tokuda, "Performance evaluation of the speaker-independent HMM-based speech synthesis system "HTS 2007" for the Blizzard Challenge 2007," in *Proc. ICASSP*, 2008, pp. 3957–3960.

[11] D. Schabus, M. Pucher, and G. Hofer, "Building a synchronous corpus of acoustic and 3D facial marker data for adaptive audio-visual speech synthesis," in *Proc. LREC*, Istanbul, Turkey, 2012, pp. 3313–3316.

[12] O. Govokhina, G. Bailly, and G. Breton, "Learning optimal audio-visual phasing for a HMM-based control model for facial animation," in *6th ISCA Workshop on Speech Synthesis (SSW6)*, Bonn, Germany, 2007, pp. 1–4.

Visual Control of Hidden-Semi-Markov-Model based Acoustic Speech Synthesis

Jakob Hollenstein[1,2]*,Michael Pucher*[1]*, Dietmar Schabus*[1,3]

[1]Telecommunications Research Center Vienna (FTW), Vienna, Austria
[2]Vienna University of Technology, Vienna, Austria
[3]Graz University of Technology, Graz, Austria
{hollenstein,pucher,schabus}@ftw.at

Abstract

We show how to visually control acoustic speech synthesis by modelling the dependency between visual and acoustic parameters within the Hidden-Semi-Markov-Model (HSMM) based speech synthesis framework. A joint audio-visual model is trained with 3D facial marker trajectories as visual features. Since the dependencies of acoustic features on visual features are only present for certain phones, we implemented a model where dependencies are estimated for a set of vowels only. A subjective evaluation consisting of a vowel identification task showed that we can transform some vowel trajectories in a phonetically meaningful way by controlling the visual parameters in PCA space. These visual parameters can also be interpreted as fundamental visual speech motion components, which leads to an intuitive control model.

Index Terms: audio-visual speech synthesis, HMM-based speech synthesis, controllability

1. Introduction

One key strength of the HSMM-based speech synthesis framework [1] lies in its greater flexibility in comparison to waveform concatenation methods, often accredited to the possibility to use model adaptation [2] and interpolation [3]. In addition to these data-driven approaches to diversify the characteristics of synthetic speech, methods that allow more direct control using phonetic background knowledge have been proposed more recently. Acoustic speech characteristics have been successfully modified by exercising control on articulartory [4] as well as on formant [5] parameters. This is achieved by training piecewise linear transformations from the models for the articulatory or formant domain to the models for the acoustic domain, using a multimodal data corpus. Similar to these works, in this paper we investigate the possibility of using visual speech features based on facial marker motion data to modify and control acoustic synthetic speech. Our work is similar to [4], but uses more restricted features (e.g., no tongue positions) which are easier to record.

Possible use cases of this include more intuitive control of speech synthesis, the possibility to use physically intuitive data to constrain trajectories as well as the possibility to use this information in language learning to provide clues of required changes.

To investigate the possibility of visual control, we modified the system used in [5] to control acoustic speech synthesis by visual features (instead of formants). The same line spectral pairs features as in [4] are adopted as acoustic features in our approach.

2. Data and System

We work with a corpus of synchronous audio and facial motion recordings [6] where facial motion was recorded using an OptiTrack system [7], which records the 3D positions of 37 markers glued to a speaker's face at 100 Hz. This corpus was originally recorded for speaker-adaptive audio-visual speech synthesis [8]. In this paper, we used one male speaker's data consisting of 223 Austrian German sentences amounting to roughly 11 minutes total.

2.1. Training

In a first attempt, we replaced the formant stream in the system described in [5] with a visual stream, resulting in a joint audio-visual model. The visual features were computed from the 3D facial marker coordinates via Principal Component Analysis (PCA) as described in [6]. However, by comparing the variation of the spectral features with respect to the different phones to the variation induced by the transformation when modifying PCA features, we found that the changes due to the transformations were very small. Hence it would be impossible to achieve a natural variety of different phones in this way. A further reduction of the number of visual dimensions lead to an even smaller amount of change induced by the transformations, hinting at the insufficient explanation of the spectral features by the visual features.

To improve the expressiveness of the features with respect to different phones and increase the ability to interpret them, PCA was abandoned and raw coordinate features used. To reduce the dimensionality of the raw visual features, a selection of some visual markers with a direct influence on speech production was made: mouth opening and lip protrusion, represented by the markers *jaw, lower lip* and *upper lip*. This is similar to the facial markers in [9]. The movement of these markers in the left/right direction was assumed to be negligible, thus only the Y and Z axes were used. A new joint audio-visual model was trained with these features. To ease visualization PCA was done on this restricted model.

It is not possible to distinguish all phones merely by their visual features. Depending on the speaker, the visual features overlap considerably for different phones, since the visual features mostly capture openness and roundedness. This explains the initial results and is similiar to the reason for using visemes in visual speech synthesis[10].

Figure 1 shows the visual features retrieved from forced alignment of the training data with respect to all vowels and diphthongs. A bagplot [11] is used to illustrate the distribution of the naturally occurring trajectories. This is done in

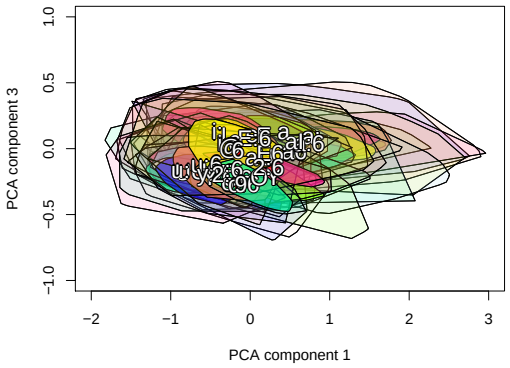


Figure 1: *Bagplot showing the distribution of vowels and diphthongs from the training data in PCA1×PCA3 space.*

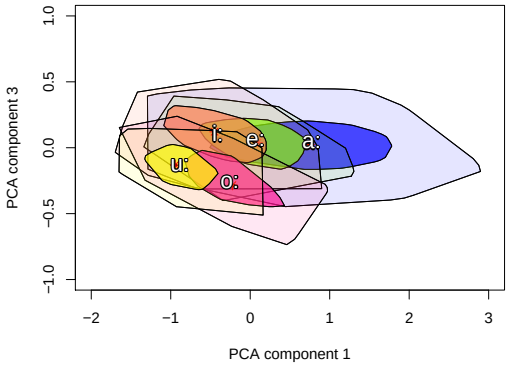


Figure 2: *Bagplot showing the distribution of selected vowels (*/a: e: i: o: u:/*) from the training data in PCA1×PCA3 space.*

PCA1×PCA3 space.

Since the visual data for different phones overlaps, we decided to constrain the dependency modeling on a small set of vowels.

The selection of phones that contribute to a transformation is a trade-off between a set comprising more and a wider variety of phones and more distinctive visual representations. For our experiments the set of vowels /a: e: i: o: u:/ was chosen. Figure 2 shows the grouping and overlap of the visual features in PCA1×PCA3 space. Notice how this resembles the vowel trapezium (rotated and mirrored, open-close from right to left). While the PCA1×PCA2 space showed even more resemblance with the vowel diagram, the PCA1×PCA3 space provides more distinction between /o: u:/ and /a: e: i:/. Which is consistent with better control regarding changes from /o: u:/ to /a: e: i:/ and vice versa and also the reason for choosing PCA1×PCA3 instead of the former.

The system uses a common clustering for acoustic and visual features, and thus for each acoustic leaf there is a corresponding visual leaf. Each leaf in the clustering tree can be assigned a transform and each transform can be assigned to several leaves. We used a single global transform for the selected vowels. To achieve this clustering, we introduced an /a: e: i: o: u:/ question at the root of the clustering tree. For all models outside of the /a: e: i: o: u:/ subtree, no transformations were trained, as illustrated in Figure 3.

Since we adapted the system from [5], we applied the same

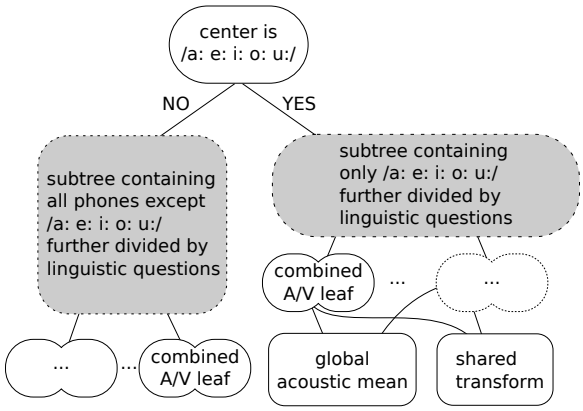


Figure 3: *Clustering tree with* /a: e: i: o: u:/ *central phone question.*

equations, described in more detail in [9], for EM estimation of the visual means $\mu_{\mathbf{Y}_j}$:

$$\mu_{\mathbf{Y}_j} = \frac{\sum_{t=1}^{T} \gamma_j(t)\mathbf{y}_t}{\sum_{t=1}^{T} \gamma_j(t)}, \tag{1}$$

where $\mathbf{y}_t$ and $\mathbf{x}_t$ describe the visual and acoustic observation vectors at time t respectively, γ_j describes the state occupancy probability of state j and T is the total length of training data.

For the acoustic means $\mu_{\hat{\mathbf{X}}_j}$ of the models in the /a: e: i: o: u:/ subtree, we use the dependency model parameters to apply a linear transformation from visual to acoustic parameters,

$$\mu_{\hat{\mathbf{X}}_j} = \frac{\sum_{t=1}^{T} \gamma_j(t)(\mathbf{x}_t - \hat{\mathbf{A}}_j\mathbf{y}_t)}{\sum_{t=1}^{T} \gamma_j(t)}. \tag{2}$$

$$\hat{\mathbf{A}}_j = \frac{\sum_{t=1}^{T} \gamma_j(t)(\mathbf{x}_t - \mu_{\hat{\mathbf{X}}_j})\mathbf{y}_t^T}{\sum_{t=1}^{T} \gamma_j(t)\mathbf{y}_t\mathbf{y}_t^T}. \tag{3}$$

The estimation of the linear transformations $\hat{\mathbf{A}}_j$ is also constrained to models in the /a: e: i: o: u:/ subtree. For the acoustic means $\mu_{\mathbf{X}_j}$ of all other phones, as well as for all (acoustic and visual) variances, we used the un-transformed version as in Equation (1).

In contrast to [9], the mean of all acoustic leafs sharing the same transform is also shared in our implementation. This is done by employing the tying mechanisms in HTS/HTK. Thus different states still share the same underlying $\hat{\mathbf{A}}_j$ regardless of the state j. The use of the tying mechanism is explained in [12].

Thus the transformation of the visual features to the audio features is not used to superimpose the modified trajectory on the original audio feature trajectory but is effectively used to generate the audio feature trajectory from the visual feature trajectory. This means that without visual information, all /a: e: i: o: u:/ phones would result in the same acoustic realization. Using this approach, we implemented a constrained audio-visual dependency modeling system.

2.2. Parameter Generation

For parameter generation, we implemented a simplified version of the algorithm described in [9]. Given an optimal state sequence, the optimal acoustic parameter sequence X_S^* is gener-

ated as

$$\mathbf{X}_S^* = (\mathbf{W}_\mathbf{X}^T \mathbf{U}_\mathbf{X}^{-1} \mathbf{W}_\mathbf{X})^{-1} \mathbf{W}_\mathbf{X}^T U_\mathbf{X}^{-1} (\mathbf{M}_\mathbf{X} + \mathbf{A}\mathbf{W}_\mathbf{Y}\mathbf{Y}_S) \quad (4)$$

which results from

$$\frac{\partial \log P(\mathbf{W}_\mathbf{X}\mathbf{X}_S, \mathbf{W}_\mathbf{Y}\mathbf{Y}_S|\lambda, q^*)}{\partial \mathbf{X}_S} = 0. \quad (5)$$

The visual parameter sequences $\mathbf{Y}_S^*$ are generated based on the approximation

$$\frac{\partial \log P(\mathbf{W}_\mathbf{Y}\mathbf{Y}_S|\lambda, q^*)}{\partial \mathbf{Y}_S} \approx \frac{\partial \log P(\mathbf{W}_\mathbf{X}\mathbf{X}_S, \mathbf{W}_\mathbf{Y}\mathbf{Y}_S|\lambda, q^*)}{\partial \mathbf{Y}_S} \quad (6)$$

which results in

$$\mathbf{Y}_S^* = (\mathbf{W}_\mathbf{Y}^T \mathbf{U}_\mathbf{Y}^{-1} \mathbf{W}_\mathbf{Y})^{-1} \mathbf{W}_\mathbf{Y}^T U_\mathbf{Y}^{-1} \mathbf{M}_\mathbf{Y} \quad (7)$$

when we set the left hand side of Equation (6) to 0. Note that this is the standard parameter generation algorithm described in [13].

3. Visual Control using PCA Features

To simplify the control from having to modify points in 6 dimensional space, we apply PCA. To modify a given model, the means are transformed into PCA space, modifications are performed relative to the resulting PCA feature vector, and the modified vector is projected back into the original space. No dimensionality reduction is used in this scheme. Also, the trajectory is not modified directly (e.g., by adding to the trajectory values), but the means are changed, thus changing the generated trajectory. This ensures smooth trajectories for the duration of the modified phone and especially at the phone boundaries. As a side effect of the smoothing, the extent of modification is slightly decreased, thus a larger change in the control parameters is required to achieve sufficiently strong effects.

The first PCA component roughly corresponds to mouth opening, while the second and third component can be interpreted as modelling rounding. Figure 4 illustrates the changes of the marker positions resulting from changes of the first PCA component between -1.5 and $+1.5$ and between -0.75 and $+0.75$ for the second PCA component.

We did not carry out a formal evaluation of the effects of the control on the visual speech motion, but synthesis of the entire 37 marker positions can be performed at the loss of some accuracy by calculating a linear regression from the 6 visual control parameters to the full visual parameter space. From examples we looked at during development, it appeared that visual synthesis is still feasible using only these parameters. There is also some loss of acoustic quality due to the simple transformation and the incomplete explanation of acoustic features by the visual features.

Figure 5 illustrates an example outcome. In the sentence *"Ich habe 'bomo' gehört"* (I heard 'bomo'), the two vowels of the nonsense word 'bomo' were modified visually by increasing and decreasing the mean of the first PCA component, corresponding to increased and decreased mouth opening, respectively. The bottom part of the figure shows the effect on the distance between the upper lip and the lower lip markers. The middle part shows the resulting spectrograms for the time segment indicated by vertical lines. The top part of the figure shows the resulting facial marker configurations at the time points indicated by small circles. Compared to the unmodified sample (± 0, center spectrogram, first formant at 287 Hz), the samples with the decreased (-1.5, left spectrogram, first formant at 408 Hz) and increased ($+1.5$, right spectrogram, first formant at 605 Hz) mouth opening exhibit a clearly visible change both in the visual trajectories as well as in the spectral power distributions.

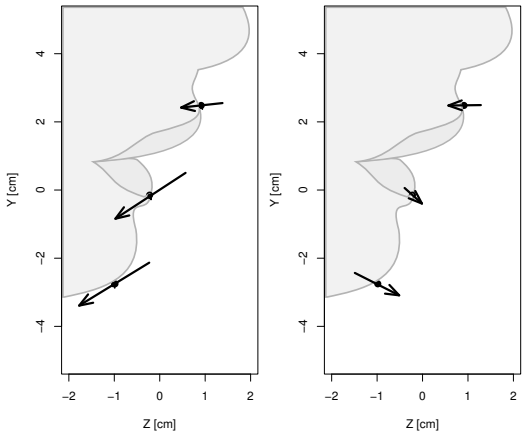


Figure 4: *Effects of changing the first PCA component from* -1.5 *to* $+1.5$ *(left) and changing the third PCA component from* -0.75 *to* $+0.75$ *(right).*

4. Evaluation

To evaluate the acoustic effects of the visual control, a subjective listening test with eleven subjects was carried out. We synthesized ten utterances containing the nonsense words 'bama', 'beme', 'bimi', 'bomo', 'bumu' and 'pata', 'pete', 'piti', 'poto', 'putu' in the carrier sentence *"Ich habe ... gehört"*, with varying visual control parameters affecting the two vowels of the nonsense word. The first PCA component was modified by applying one of the three offsets $-1.5, \pm 0, +1.5$ and the third PCA component by applying one of the three offsets $-0.75, \pm 0, +0.75$, resulting in nine different realizations[1].

Each test subject heard all 90 synthesized examples in random order and was asked to identify what they heard as one of the five variants of the nonsense word (either 'bama', 'beme', 'bimi', 'bomo', 'bumu' or 'pata', 'pete', 'piti', 'poto', 'putu') or none of these. The results are given in Table 1, where the first column gives the original vowel, the second and third column give the applied offsets for the first and third PCA components, and the remaining columns give the identification percentages. These results are also visualized in Figure 6 as stacked bar charts. For each initial vowel, the central bar corresponds to the unmodified sample, the upper-left bar corresponds to a shift in the upper-left direction (compare Figure 2), etc.

In most cases, there is a clear majority regarding the perceived vowel and in these cases the change is consistent with what can be expected when we consider the vowel distributions of Figure 2. For each initial vowel, we can successfully transform towards at least one other vowel. It is interesting that for each of them there is a direction in which the listeners perceived the original vowel more clearly, i.e. in a higher percentage of cases. In all five cases, this direction is leading "away" from the other vowels (Figure 2). Furthermore, we see that for /a/, there is a fairly large number of "no match" votes ("?") in all nine cases. Part of this is due to acoustic artifacts (e.g., buzzing

[1] Examples on http://userver.ftw.at/~schabus/avsp2013vc

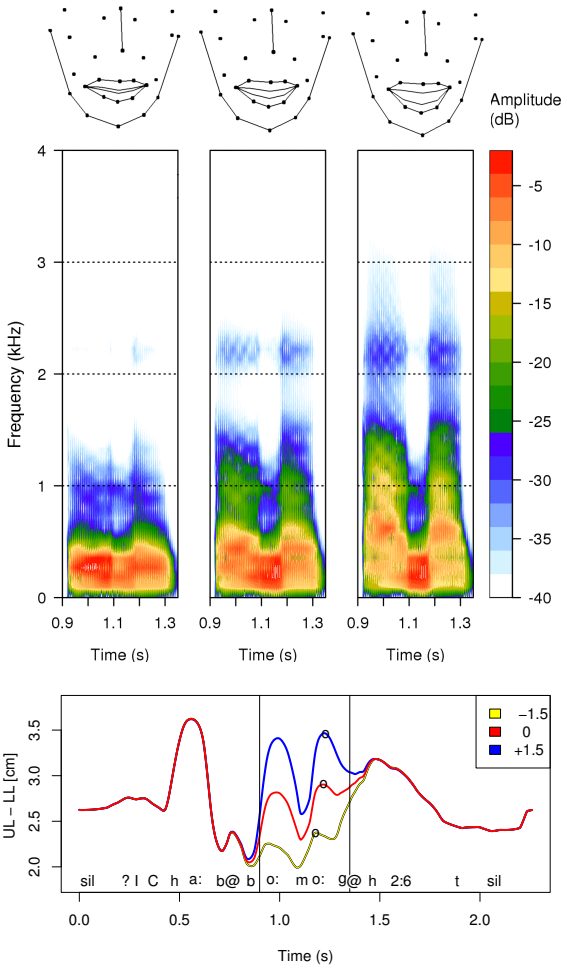


Figure 5: *Example outcome of modification of the first PCA component. Bottom: Distance between the upper lip and lower lip markers over time. Middle: spectrograms for the time segment indicated by vertical lines. Top: Facial marker configuration at the time points indicated by small circles.*

or distortions regarding amplitude). These artifacts can be attributed to leaving the area of the PCA space in which observations are naturally occurring and thus creating artificial visual features and inducing artificial sound.

In some examples (e.g., the one in Figure 5) we saw that the closure between the two modified vowels was not synthesized correctly in the visual domain as the smooth trajectory generation prevented the lip movement from reaching the closure point. This could be prevented, e.g., by decreasing the variances and thus forcing the trajectories closer to the specified means, or by modifying the dynamic features. This also indicates that PCA transformation alone does not capture all possible modifications of the underlying feature space adequately and a different parametrization for exercising control may be necessary.

We saw in our experiments that the offsets applied in PCA space needed to be larger than expected (i.e., 1.5 and 0.75 instead of 1.0 and 0.5) in order to properly induce changes. We attribute part of this to the generation algorithm producing over-

Table 1: Evaluation Results: Identification percentages for each initial vowel, modified by each of nine control offset combinations.

V	Δ_1	Δ_3	a	e	i	o	u	?
a	-1.5	+0.75	0	77.3	0	0	0	22.7
a	0	+0.75	4.5	72.7	0	0	0	22.7
a	+1.5	+0.75	9.1	40.9	0	0	0	50.0
a	-1.5	0	22.7	36.4	0	0	0	40.9
a	0	0	63.6	0	0	0	0	36.4
a	+1.5	0	72.7	0	0	0	0	27.3
a	-1.5	-0.75	0	0	0	68.2	0	31.8
a	0	-0.75	4.5	0	0	54.5	0	40.9
a	+1.5	-0.75	36.4	0	0	18.2	0	45.5
e	-1.5	+0.75	0	27.3	72.7	0	0	0
e	0	+0.75	0	100.0	0	0	0	0
e	+1.5	+0.75	0	100.0	0	0	0	0
e	-1.5	0	0	0	90.9	0	9.1	0
e	0	0	0	68.2	22.7	0	0	9.1
e	+1.5	0	4.5	95.5	0	0	0	0
e	-1.5	-0.75	0	0	45.5	0	50.0	4.5
e	0	-0.75	0	0	13.6	0	72.7	13.6
e	+1.5	-0.75	13.6	31.8	0	31.8	18.2	4.5
i	-1.5	+0.75	0	4.5	86.4	0	9.1	0
i	0	+0.75	0	95.5	4.5	0	0	0
i	+1.5	+0.75	0	95.5	4.5	0	0	0
i	-1.5	0	0	0	95.5	0	4.5	0
i	0	0	0	31.8	50.0	4.5	13.6	0
i	+1.5	0	0	81.8	4.5	9.1	4.5	0
i	-1.5	-0.75	0	0	54.5	0	40.9	4.5
i	0	-0.75	0	0	40.9	0	50.0	9.1
i	+1.5	-0.75	0	50.0	0	27.3	18.2	4.5
o	-1.5	+0.75	0	63.6	13.6	0	13.6	9.1
o	0	+0.75	0	86.4	0	0	0	13.6
o	+1.5	+0.75	4.5	86.4	0	0	0	9.1
o	-1.5	0	0	0	4.5	0	90.9	4.5
o	0	0	0	18.2	0	72.7	9.1	0
o	+1.5	0	72.7	9.1	0	13.6	0	4.5
o	-1.5	-0.75	0	0	0	0	100.0	0
o	0	-0.75	0	0	0	27.3	72.7	0
o	+1.5	-0.75	4.5	0	0	90.9	0	4.5
u	-1.5	+0.75	0	0	95.5	0	4.5	0
u	0	+0.75	0	22.7	68.2	0	0	9.1
u	+1.5	+0.75	0	81.8	13.6	0	0	4.5
u	-1.5	0	0	0	27.3	0	72.7	0
u	0	0	0	0	22.7	0	63.6	13.6
u	+1.5	0	0	27.3	4.5	9.1	50.0	9.1
u	-1.5	-0.75	0	0	0	0	90.9	9.1
u	0	-0.75	0	0	0	0	90.9	9.1
u	+1.5	-0.75	0	0	0	0	95.5	4.5

smoothed trajectories and part of it to the overlapping which essentially requires us to leave ambiguous areas to create unambiguous sounds.

5. Conclusion

In this paper we have shown that – similar to previous work, where acoustic speech synthesis is controlled via articulatory or formant features – it is also possible to achieve phonetically meaningful transformations of acoustic synthetic speech by exercising control in terms of visual speech features, namely 3D facial marker motion data. This can be seen as a more restrictive setting than 3D articulatory data, because we have no information on the position of the tongue.

For all of the selected five phones, transformations to at least one other phone have been shown to be feasible, as determined by a subjective listening test with eleven subjects. The acoustic phone realizations resulting from changes in certain directions are consistent with the distribution in visual PCA space.

Future work could explore the improvements achievable by using a more sophisticated control model for example using combined per-state and per-context transformations. Some improvement may also be possible by using more descriptive

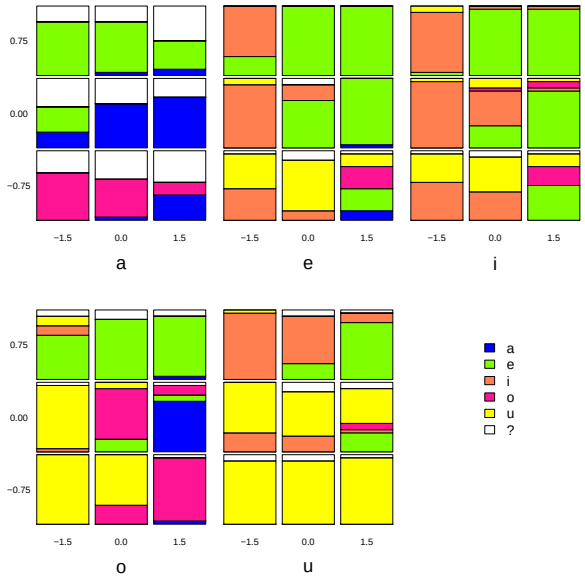


Figure 6: *Visualization of the evaluation results (Table 1). For each initial phone, the central subplot shows the classification results for the unmodified phone. The eight surrounding subplots show the classification results for the modified phones. Colors and orientation are in line with Figure 2.*

features, rather than only three markers on the lips and jaw. Additionally, we would like to evaluate the coherence of the modified visual and acoustic speech signals in combined audio-visual perceptive experiments. Another interesting topic to investigate would be to combine both facial marker and articulatory data, requiring a synchronous multimodal corpus, which we plan to build in the near future.

6. Acknowledgements

We want to thank Korin Richmond for providing a HMM-based system with dependency modeling. This work was supported by the Austrian Science Fund (FWF): P22890-N23. The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH is funded within the program COMET - Competence Centers for Excellent Technologies by BMVIT, BMWA, and the City of Vienna. The COMET program is managed by the FFG.

7. References

[1] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. SSW*, 2007, pp. 294–299.

[2] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[3] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," in *Proc. Eurospeech*, 1997.

[4] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge," in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008, pp. 573–576.

[5] M. Lei, J. Yamagishi, K. Richmond, Z.-H. Ling, S. King, and L.-R. Dai, "Formant-controlled HMM-based speech synthesis," in *Proc. Interspeech*, Florence, Italy, August 2011, pp. 2777–2780.

[6] D. Schabus, M. Pucher, and G. Hofer, "Building a synchronous corpus of acoustic and 3D facial marker data for adaptive audio-visual speech synthesis," in *Proc. LREC*, Istanbul, Turkey, May 2012, pp. 3313–3316.

[7] Naturalpoint, 2013. [Online]. Available: http://www.naturalpoint.com/optitrack/

[8] D. Schabus, M. Pucher, and G. Hofer, "Speaker-adaptive visual speech synthesis in the HMM-framework," in *Proc. Interspeech*, Portland, OR, USA, 2012, pp. 979–982.

[9] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *Trans. Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, 2009.

[10] T. Chen, "Audiovisual speech processing," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 9–21, 2001.

[11] P. J. Rousseeuw, I. Ruts, and J. W. Tukey, "The bagplot: a bivariate boxplot," *The American Statistician*, vol. 53, no. 4, pp. 382–387, 1999.

[12] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[13] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP*, Detroit, MI, USA, 1995, pp. 660–663.

2.4 Speaker Verification Spoofing

The general synthetic signal impostor problem can be summarized as follows:

- Given a classification system C that can classify (verify / identify) signals $x_1, .., x_T$ of a certain type A.
- We can build a regression system R that can generate synthetic signals $x_{1_S}, .., x_{T_S}$ of type A_S.
- Given that both systems C and R have a similar state-of-the art performance, it is likely that $x_{1_S}, .., x_{T_S}$ are correctly classified by C.
- For security we would need a system that classifies signals and verifies that they are of the appropriate type A, which includes that they were produced by a certain source etc.

The objective in speaker verification (SV) is to accept or reject a claim of identity based on a voice sample Reynolds et al. [2000a]. Many investigations on the imposture problem as related to SV have been reported over the years as well as methods to prevent such impostures. The simplest imposture is playback of a voice recording for a targeted speaker and the well-known solution is a text-prompted approach Matsui and Furui [1995]. In addition, the vulnerability of SV to voice mimicking by humans has also been examined in Lau et al. [2004]; Sullivan and Pelecanos [2001]. On the other hand, advanced speech technologies present new problems for SV systems including imposture using speech manipulation of a recorded voice via analysis-by-resynthesis methods Genoud and Chollet [1998]; Pellom and Hansen [1999]; Lindberg and Blomberg [1999], voice conversion of the recorded voice Matrouf et al. [2006]; Bonastre et al. [2007, 2006]; Farrus et al. [2008], and diphone speech synthesis methods Lindberg and Blomberg [1999].

The use of synthesized speech potentially poses two related problems for SV systems. The first problem is confirmation of an acquired speech signal as having originated from a particular individual. In this case, the speech signal might be incorrectly confirmed as having originated from an individual when in fact the speech signal is synthetic. The second problem is in remote or on-line authentication where voice is used. In this case, a synthesized speech signal could be used to wrongly gain access to a person's account and text-prompting would not present a problem for a text-to-speech (TTS) system. In both of these problems, the speech model for the synthesizer must be targeted to a specific person's voice. SV is also being used in forensic applications Bo [2000] and therefore security against imposture is also of obvious importance.

The problem of imposture against SV systems using synthetic speech was first published over 10 years ago by Masuko, et al. Masuko et al. [1999]. In their original work, the authors used a hidden Markov model (HMM)-based text-prompted SV system Matsui and Furui [1995] and an HMM-based TTS synthesizer. In the SV system, feature vectors were scored against speaker and background models composed of concatenated phoneme

models. The acoustic models used in the speech synthesizer were adapted to each of the human speakers Masuko et al. [1996, 1997]. When tested with 20 human speakers, the system had a 0% false acceptance rate (FAR) and 7.2% false rejection rate (FRR); when tested with synthetic speech, the system accepted over 70% of matched claims, i.e. a synthetic signal matched to a targeted speaker and an identity claim of that same speaker.

In subsequent work by Masuko, et al. Masuko et al. [2000], the authors extended the research in two ways. First, they improved their synthesizer by generating speech using F_0 (fundamental frequency) information. Second, they improved their SV system by utilizing both F_0 and spectral information. The F_0 modeling techniques used in synthesis were the same used in the SV system. By improving the SV system, the authors were able to lower the matched claim rate for synthetic speech to 32%, however, the FAR for the human speech increased to 1.8%.

In the last 10 years, both SV and TTS systems have improved dramatically. Around the same time as Masuko's work, Gaussian mixture model-universal background model (GMM-UBM) SV systems were first proposed Reynolds et al. [2000a]. Since this time, GMM-UBM based SV systems have produced excellent performance and have achieved equal error rates (EERs) of 0.1% on the TIMIT corpus (ideal recordings) and 12% on NIST 2002 Speaker Recognition Evaluations (SRE) (non-ideal recordings) Bimbot et al. [2004]; Kinnunen et al. [2006]. Newer systems based on support vector machines (SVMs) using GMM supervectors have been proposed and in some cases can lead to lower EERs Campbell et al. [2006], Longworth and Gales [2009].

Until recently, developing a TTS synthesizer for a targeted speaker required a large amount of speech data from a carefully prepared transcript in order to construct the speech model. However, with a state-of-the-art HMM-based TTS synthesizer Zen et al. [2009b], the speech model can now be adapted from an average model (derived from other speakers) or a background model (derived from one speaker) using only a small amount of speech data. Moreover, recent experiments with HMM-based speech synthesis systems have also demonstrated that the speaker-adaptive HMM-based speech synthesis is robust to non-ideal speech data that are recorded under various conditions and with varying microphones, that are not perfectly clean, and/or that lack phonetic balance. In Yamagishi et al. [2009e] a high-quality voice was built from audio collected off of the Internet. This data was not recorded in a studio, had a small amount of background noise, and the microphones varied in the data. Further Yamagishi et al. [2010] reported construction of thousands of voices for HMM-based speech synthesis based on corpora such as the Wall Street Journal (WSJ0, WSJ1, and WSJCAM0), Resource Management, Globalphone and SPEECON. Taken together, these state-of-the-art speech synthesizers pose new challenges to SV systems.

In prior work, we utilized a state-of-the-art TTS synthesizer and revisited the problem of imposture using a GMM-UBM SV system with a small speech corpus [De Leon et al., 2010a] and then extended to a larger corpus [De Leon et al., 2010b]. Recently,

we examined the performance using the SVM-based SV system and initial experiments on detecting a synthetic speech signal [De Leon et al., 2011]. In our work, we provide complete evaluations using both GMM-UBM and SVM-based SV systems and new results from a proposed synthetic speech detector (SSD) which uses phase-based features for classification. First, we train two different SV systems (GMM-UBM and SVM using GMM supervectors) using human speech (283 speakers from the WSJ corpus). Second, we create synthetic test speech for each of the 283 speakers by adapting a background model to the targeted speaker. Finally, we measure EER and true acceptance rates when tested using human speech and measure the matched claim rate using synthetic speech. As we will demonstrate, the matched claim rate is above 81% for each of the SV systems hence the vulnerability of the SV systems to synthetic speech. Next, we turn our attention to detection of synthetic speech as a means to prevent imposture by synthetic speech. We summarize results with a previously-proposed method which uses average inter-frame difference of log-likelihood (IFDLL) [Satoh et al., 2001] and show that this is no longer a viable discriminator for high-quality synthetic speech such as that which we are using. Instead, we propose a new discrimination feature based on relative phase shift (RPS) and show that this can be used to reliably detect synthetic speech. We also show a simple and effective method for training the classifier using transcoded human speech as a surrogate for synthetic speech.

2280 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 8, OCTOBER 2012

Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech

Phillip L. De Leon, *Senior Member, IEEE*, Michael Pucher, *Member, IEEE*, Junichi Yamagishi, Inma Hernaez, and Ibon Saratxaga

***Abstract*—In this paper, we evaluate the vulnerability of speaker verification (SV) systems to synthetic speech. The SV systems are based on either the Gaussian mixture model–universal background model (GMM-UBM) or support vector machine (SVM) using GMM supervectors. We use a hidden Markov model (HMM)-based text-to-speech (TTS) synthesizer, which can synthesize speech for a target speaker using small amounts of training data through model adaptation of an average voice or background model. Although the SV systems have a very low equal error rate (EER), when tested with synthetic speech generated from speaker models derived from the Wall Street Journal (WSJ) speech corpus, over 81% of the matched claims are accepted. This result suggests vulnerability in SV systems and thus a need to accurately detect synthetic speech. We propose a new feature based on relative phase shift (RPS), demonstrate reliable detection of synthetic speech, and show how this classifier can be used to improve security of SV systems.**

***Index Terms*—Security, speaker recognition, speech synthesis.**

I. Introduction

THE objective in speaker verification (SV) is to accept or reject a claim of identity based on a voice sample [1]. Many investigations on the imposture problem as related to SV have been reported over the years as well as methods to prevent such impostures. The simplest imposture is playback of a voice recording for a targeted speaker and the well-known solution is a text-prompted approach [2]. In addition, the vulnerability of SV to voice mimicking by humans has also been examined in [3], [4]. On the other hand, advanced speech technologies present new problems for SV systems including imposture using speech manipulation of a recorded voice via analysis-by-resynthesis methods [5]–[7], voice conversion of the recorded voice [8]–[11], and diphone speech synthesis methods [7].

The use of synthesized speech potentially poses two related problems for SV systems. The first problem is confirmation of an acquired speech signal as having originated from a particular individual. In this case, the speech signal might be incorrectly confirmed as having originated from an individual when in fact the speech signal is synthetic. The second problem is in remote or online authentication where voice is used. In this case, a synthesized speech signal could be used to wrongly gain access to a person's account and text-prompting would not present a problem for a text-to-speech (TTS) system. In both of these problems, the speech model for the synthesizer must be targeted to a specific person's voice. SV is also being used in forensic applications [12] and therefore security against imposture is also of obvious importance.

The problem of imposture against SV systems using synthetic speech was first published over 10 years ago by Masuko, *et al.* [13]. In their original work, the authors used a hidden Markov model (HMM)-based text-prompted SV system [2] and an HMM-based TTS synthesizer. In the SV system, feature vectors were scored against speaker and background models composed of concatenated phoneme models. The acoustic models used in the speech synthesizer were adapted to each of the human speakers [14], [15]. When tested with 20 human speakers, the system had a 0% false acceptance rate (FAR) and 7.2% false rejection rate (FRR); when tested with synthetic speech, the system accepted over 70% of matched claims, i.e. a synthetic signal matched to a targeted speaker and an identity claim of that same speaker.

In subsequent work by Masuko, *et al.* [16], the authors extended the research in two ways. First, they improved their synthesizer by generating speech using F_0 (fundamental frequency) information. Second, they improved their SV system by utilizing both F_0 and spectral information. The F_0 modeling techniques used in synthesis were the same used in the SV system. By improving the SV system, the authors were able to lower the matched claim rate for synthetic speech to 32%, however, the FAR for the human speech increased to 1.8%.

In the last 10 years, both SV and TTS systems have improved dramatically. Around the same time as Masuko's

Manuscript received May 20, 2011; revised November 22, 2011 and April 20, 2012; accepted May 16, 2012. Date of publication May 25, 2012; date of current version August 13, 2012. This work was supported in part by the Austrian Science Fund (FWF): P22890-N23. The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH was supported within the program COMET Competence Centers for Excellent Technologies by BMVIT, BMWA, and the City of Vienna. The COMET program is managed by the FFG. The work of J. Yamagishi was supported by the Engineering and Physical Sciences Research Council (EPSRC) under Grants EP/I031022/1 and EP/J002526/1. This work was supported in part by the Spanish Government (TEC2009-14094-C04-02) and the Basque Government (IE09-262, MV20090225). This work was presented in part at the 2010 and 2011 International Conference on Acoustics, Speech, and Signal Processing. (ICASSP) and at the 2010 Odyssey Speaker Recognition Workshop. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nestor Becerra Yoma.

P. L. De Leon is with the Klipsch School of Electrical and Computer Engineering, New Mexico State University (NMSU), Las Cruces, NM 88003 USA (e-mail: pdeleon@nmsu.edu).

M. Pucher is with the Telecommunications Research Center Vienna (FTW), 1220 Vienna, Austria (e-mail: pucher@ftw.at).

J. Yamagishi is with the University of Edinburgh, Edinburgh, EH8 9AB, U.K. (e-mail: jyamagis@inf.ed.ac.uk).

I. Hernaez and I. Saratxaga are with University of the Basque Country, Bilbao 48013, Spain (e-mail: inma@aholab.ehu.es; ibon@aholab.ehu.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2201472

work, Gaussian mixture model-universal background model (GMM-UBM) SV systems were first proposed [1]. Since this time, GMM-UBM based SV systems have produced excellent performance and have achieved equal error rates (EERs) of 0.1% on the TIMIT corpus (ideal recordings) and 12% on NIST 2002 Speaker Recognition Evaluations (SRE) (non-ideal recordings) [17], [18]. Newer systems based on support vector machines (SVMs) using GMM supervectors have been proposed and in some cases can lead to lower EERs [19], [20].

Until recently, developing a TTS synthesizer for a targeted speaker required a large amount of speech data from a carefully prepared transcript in order to construct the speech model. However, with a state-of-the-art HMM-based TTS synthesizer [21], the speech model can now be adapted from an average model (derived from other speakers) or a background model (derived from one speaker) using only a small amount of speech data. Moreover, recent experiments with HMM-based speech synthesis systems have also demonstrated that the speaker-adaptive HMM-based speech synthesis is robust to non-ideal speech data that are recorded under various conditions and with varying microphones, that are not perfectly clean, and/or that lack phonetic balance. In [22] a high-quality voice was built from audio collected off of the Internet. This data was not recorded in a studio, had a small amount of background noise, and the microphones varied in the data. Further [23] reported construction of thousands of voices for HMM-based speech synthesis based on corpora such as the Wall Street Journal (WSJ0, WSJ1, and WSJCAM0), Resource Management, Globalphone, and SPEECON. Taken together, these state-of-the-art speech synthesizers pose new challenges to SV systems.

In prior work, we utilized a state-of-the-art TTS synthesizer and revisited the problem of imposture using a GMM-UBM SV system with a small speech corpus [24] and then extended to a larger corpus [25]. Recently, we examined the performance using the SVM-based SV system and initial experiments on detecting a synthetic speech signal [26]. In this paper, we provide complete evaluations using both GMM-UBM and SVM-based SV systems and new results from a proposed synthetic speech detector (SSD) which uses phase-based features for classification. First, we train two different SV systems (GMM-UBM and SVM using GMM supervectors) using human speech (283 speakers from the WSJ corpus). Second, we create synthetic test speech for each of the 283 speakers by adapting a background model to the targeted speaker. Finally, we measure EER and true acceptance rates when tested using human speech and measure the matched claim rate using synthetic speech. As we will demonstrate, the matched claim rate is above 81% for each of the SV systems hence the vulnerability of the SV systems to synthetic speech. Next, we turn our attention to detection of synthetic speech as a means to prevent imposture by synthetic speech. We summarize results with a previously proposed method which uses average inter-frame difference of log-likelihood (IFDLL) [27] and show that this is no longer a viable discriminator for high-quality synthetic speech such as that which we are using. Instead, we propose a new discrimination feature based on relative phase shift (RPS) and show that this can be used to reliably detect synthetic speech. We also show a simple and effective method for training the classifier using transcoded human speech as a surrogate for synthetic speech.

This paper is organized as follows. In Sections II and III, we provide brief overviews of the SV and TTS systems. In Section IV, we review IFDLL and provide details on our proposed RPS feature for detecting synthetic speech. In Section V, we describe the WSJ corpus and explain how we partitioned the corpus for training and testing of all the required systems. We note that although the WSJ corpus is not a standard corpus for SV research, it is one of the few that provides sufficient speech material from hundreds of speakers which is required to construct synthetic voices matched to their human counterparts. Section VI gives the evaluation results using the WSJ corpus and its synthesized counterpart as well as the results when using RPS to detect synthetic speech. Finally, we conclude the article in Section VII.

II. Speaker Verification Systems

Our SV systems are based on the well-known GMM-UBM described in [17] and the SVM using GMM supervectors described in [19]. We briefly review these systems and our implementation in the following subsections.

A. SV System Training

For both SV systems, T feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ are extracted every 10 ms using a 25-ms hamming window and composed of 15 mel-frequency cepstral coefficients (MFCCs), 15 delta MFCCs, log energy, and delta-log energy as elements. We apply feature warping to the vectors in order to improve robustness [28] which is adequate given the high-quality recordings in the WSJ corpus.

Training the GMM-UBM system is composed of two stages, shown in Fig. 1(a) and (b). The SVM using GMM supervectors system includes these two stages and two additional stages shown in Fig. 1(c) and (d). In the first stage, a GMM-UBM consisting of the model parameters $\lambda_{\text{UBM}} = \{w_i, \boldsymbol{\eta}_i, \boldsymbol{\Sigma}_i\}$ is constructed from the collection of speakers' feature vectors. Here, we assume $M = 512$ component densities in the GMM-UBM and w_i, $\boldsymbol{\eta}_i$, and $\boldsymbol{\Sigma}_i$ represent, respectively, the weight, mean vector, and diagonal covariance matrix of the ith component density where $1 \leq i \leq M$. These parameters are estimated using the expectation maximization (EM) algorithm. In practice the GMM-UBM is constructed from non-target speakers.

In the second stage, feature vectors are extracted from target speakers' utterances. We assume the availability of several utterances per speaker recorded (preferably) under different channel conditions in order to improve the speaker modeling and robustness of the system. Feature vectors from each utterance are used to maximum *a posteriori* (MAP)-adapt only the mean vectors of the GMM-UBM to form speaker- and utterance-dependent models $\lambda_{s,u} = \{w_i, \boldsymbol{\mu}_{s,u,i}, \boldsymbol{\Sigma}_i\}$ where $\boldsymbol{\mu}_{s,u,i}$ is the MAP-adapted mean vector of the ith component density from speaker s and utterance u.

In the third stage (used for the SVM), the mean vectors $\boldsymbol{\mu}_{s,u,i}$ are then diagonally scaled according to

$$\mathbf{m}_{s,u,i} = \sqrt{w_i}\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\mu}_{s,u,i} \tag{1}$$

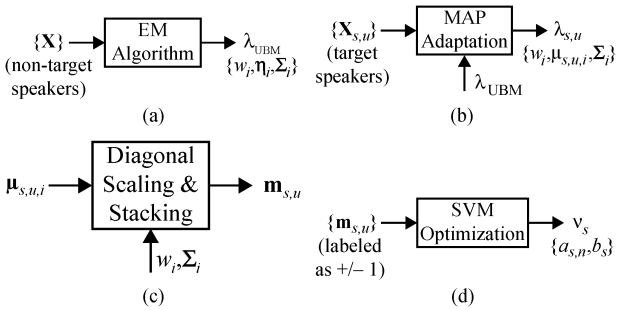


Fig. 1. Stages of training the SV systems. The GMM-UBM SV system is trained with (a)-(b) and the SVM SV system is trained with (a)–(d). Although the GMM-UBM is normally derived from non-target speakers, as described in Section V, we have used target speakers.

and stacked to form a GMM supervector for a speaker's given utterance

$$\mathbf{m}_{s,u} = \begin{bmatrix} \mathbf{m}_{s,u,1} \\ \vdots \\ \mathbf{m}_{s,u,M} \end{bmatrix}. \tag{2}$$

In the fourth stage (used for the SVM), the target speaker's supervectors are labeled as $+1$ and all other speakers' supervectors as -1. Parameters (weights, a_n and bias, b) of the SVM using a linear kernel are computed for each speaker through an optimization process. As derived in [29], an appropriately chosen distance measure between the mean vectors $\boldsymbol{\mu}_{s,u,i}$, results in a corresponding linear kernel involving the supervectors in (2) composed of diagonally scaled mean vectors (1).

In conventional GMM-UBM SV systems, we normally assume a single training signal (or several utterances concatenated to form a single training signal) so that the sth speaker model is simply $\lambda_s = \{w_i, \mu_{s,i}, \Sigma_i\}$. For the SVM, the speaker model is denoted $\nu_s = \{a_{s,n}, b_s\}$ where $a_{s,n}$ is the weight of the nth support vector, b_s is the bias, and $n \in \mathcal{S}$ and $\mathcal{S}$ is the set of indices of the support vectors.

B. SV System Testing

In SV system testing we are given an identity claim C and feature vectors $\mathbf{X}$ from a test utterance and must accept or reject the claim. For the GMM-UBM system, we compute the log-likelihood ratio

$$\Lambda(\mathbf{X}) = \log p(\mathbf{X}|\lambda_C) - \log p(\mathbf{X}|\lambda_{\mathrm{UBM}}). \tag{3}$$

where

$$\log p(\mathbf{X}|\lambda) = \frac{1}{R}\sum_{n=1}^{R} \log p(\mathbf{x}_n|\lambda) \tag{4}$$

and R is the number of test feature vectors. The claimant speaker is accepted if

$$\Lambda(\mathbf{X}) \geq \theta \tag{5}$$

where θ is the decision threshold. In the SVM system, the supervector $\mathbf{m}_{\mathrm{test}}$ is computed from the feature vectors $\mathbf{X}$ by essentially repeating stages 2 and 3 from training. We then compute

$$y(\mathbf{X}) = \sum_{n\in\mathcal{S}} a_{C,n} l_{C,n} \mathbf{m}_{\mathrm{test}}^T \mathbf{m}_{C,n} + b_C \tag{6}$$

where $l_{C,n}$ denotes the labels associated with the support vectors and accept the claim if $y(\mathbf{X}) \geq 0$.

III. Text-to-Speech Synthesizer

Our TTS systems are based on the well-known statistical parametric speech synthesis framework described in [21]. The speaker adaptation techniques of the framework allows us to generate a personalized synthetic voice using as little as a few minutes of recorded speech from a target speaker and we use the techniques for building the personalized synthetic voices for hundreds of speakers.[1] In the following subsections, we briefly review our TTS systems and our implementation.

A. TTS System Training

Our TTS system is built using the framework from the "HTS-2008" system [22], which was a speaker-adaptive system entered for the Blizzard Challenge 2008 [31]. In the challenge, the system had the equal best naturalness and the equal best intelligibility on a training data set comprising one hour of speech. The system was also found to be as intelligible as human speech [32]. The speech synthesis system consists of three main components: speech analysis and average voice training, speaker adaptation, and speech generation.

In the speech analysis and the average voice training component, three kinds of parameters for the STRAIGHT (Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram [33]) mel-cepstral vocoder with mixed excitation (i.e., 39-dimensional mel-cepstral coefficients, $\log F_0$ and five-dimensional band-limited aperiodicity measures) are extracted as feature vectors for HMMs [34]. Context-dependent, multi-stream, left-to-right, multi-space distribution (MSD), hidden semi-Markov models (HSMMs) [35] are trained on multi-speaker databases in order to simultaneously model the acoustic features and duration. A set of model parameters (Gaussian mean vectors and diagonal covariance matrices) for the speaker-independent MSD-HSMMs are estimated using the EM algorithm. First, speaker-independent monophone MSD-HSMMs are trained from an initial segmentation, converted into context-dependent MSD-HSMMs, and reestimated. Then, decision-tree-based context clustering with

[1]We are not considering unit selection and concatenative speech synthesis which is used in some commercial speech synthesizers [30]. Developing the unit selection and concatenation synthesizer for a targeted speaker requires a large amount of speech data, at least one hour, from a carefully prepared transcript. Therefore, we believe this approach is unlikely to be used, in practice, for imposture against SV systems in contrast to HMM-based TTS systems, which requires much smaller amounts of speech. It is possible, however, to use "voice conversion" techniques to change the speaker in the unit selection synthesizer and there are reports [8]–[11] of this approach being used for imposture against SV systems. We note that voice conversion approaches use similar vocoders to statistical parametric speech synthesis and we hypothesize that the proposed synthetic speech detection method would also be effective with voice conversions.

the minimum description length (MDL) criterion [36] is applied to the HSMMs and the model parameters of the HSMMs are tied at leaf nodes. The clustered HSMMs are reestimated again. The clustering processes are repeated twice and the whole process is further repeated twice using segmentation labels refined with the trained models in a bootstrap manner. All re-estimation and resegmentation processes utilize speaker-adaptive training (SAT) [37] based on constrained maximum-likelihood linear regression (CMLLR) [38].

B. TTS System Adaptation

In the speaker adaptation component, the speaker-independent MSD-HSMMs are transformed by using constrained structural maximum *a posteriori* linear regression (CSMAPLR) [39]. Note that not only output pdfs for the acoustic features but also duration models are transformed in the speaker adaptation. This adaptation requires as little as a few minutes of recorded speech from a target speaker in order to generate a personalized synthetic voice.

C. TTS System Synthesis

In the speech generation component, acoustic feature parameters are generated from the adapted MSD-HSMMs using a parameter generation algorithm that considers both the global variance of a trajectory to be generated and trajectory likelihood [40]. Finally an excitation signal is generated using mixed excitation (pulse plus band-filtered noise components) and pitch-synchronous overlap and add (PSOLA) [41]. This signal is used to excite a mel-logarithmic spectrum approximation (MLSA) filter [42] corresponding to the STRAIGHT mel-cepstral coefficients to generate the synthetic speech waveform.

IV. Detection of Synthetic Speech

In this section, we begin by evaluating the average IFDLL, previously proposed in [27] to detect synthetic speech. As we demonstrate, average IFDLL is no longer a viable discriminator for state-of-the-art HMM-based synthetic speech such as that which we are using. Based on these results, we then propose a more accurate GMM-based classifier based on the RPS feature. The use of a phase-based feature extracted directly from the speech signal is a novel application in the detection of synthetic speech.

A. Average Inter-Frame Difference of Log-Likelihood

The IFDLL is defined as [27]

$$\Delta_n = |\log p(\mathbf{x}_n|\lambda_C) - \log p(\mathbf{x}_{n-1}|\lambda_C)| \tag{7}$$

and the average IFDLL is given by

$$\bar{\Delta} = \frac{1}{R}\sum_{n=1}^{R}\Delta_n. \tag{8}$$

The authors in [27] observed that for synthetic speech, average IFDLL is significantly lower than that for human speech and can be used as a discriminator. This difference was explained as a result of the HMM-based synthesizer, used in the work, generating a speech parameter sequence so as to maximize the output probability. This maximization normally leads to a time variation of the speech parameters of synthetic speech becoming smaller than that for human speech.

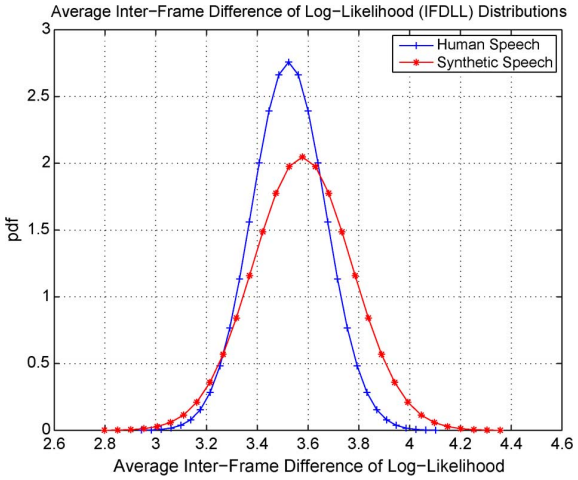


Fig. 2. Approximate distributions of average interframe-difference of log-likelihood for human and synthetic speech. Due to the overlapping distributions, the average IFDLL cannot be used to detect synthetic speech.

In Fig. 2, we show the approximate distributions of average IFDLL for human and synthetic speech using the 283 speaker WSJ corpus (subsets HS-B and TTS-B as described in Section V). Using the state-of-the-art HMM-based speech synthesizer described in Section III, this measure no longer appears to be robust enough to detect synthetic speech, since the distributions of average IFDLL for human and synthetic speech have significant overlap. In [25], we also showed that dynamic-time-warping of MFCC features and automatic speech recognition (ASR) word-error-rate are also not robust measures to detect synthetic speech.

B. Relative Phase Shift

Since the human auditory system is known to be relatively insensitive to the speech signal's phase [43], the vocoder used in TTS is normally based on a minimum-phase vocal tract model for simplicity. This simplification leads to differences in the phase spectra between human and synthetic speech which are not usually audible. However, these differences can be used to construct a new feature which allows detection of synthetic speech.

We propose using the RPS representation of the harmonic phase as a discriminating feature for detecting synthetic speech. The RPS is described in [44], [45] and is based on the harmonic modeling of the speech signal [46]. In these models, the harmonic part of the speech signal may be represented as

$$h(t) = \sum_k A_k(t)\cos\left[\Phi_k(t)\right] \tag{9}$$

where $A_k(t)$ is the amplitude and

$$\Phi_k(t) = 2\pi F_0 kt + \theta_k \tag{10}$$

is the instantaneous phase of the kth harmonic. Here we denote the initial phase of the kth harmonic as θ_k. The RPS values for

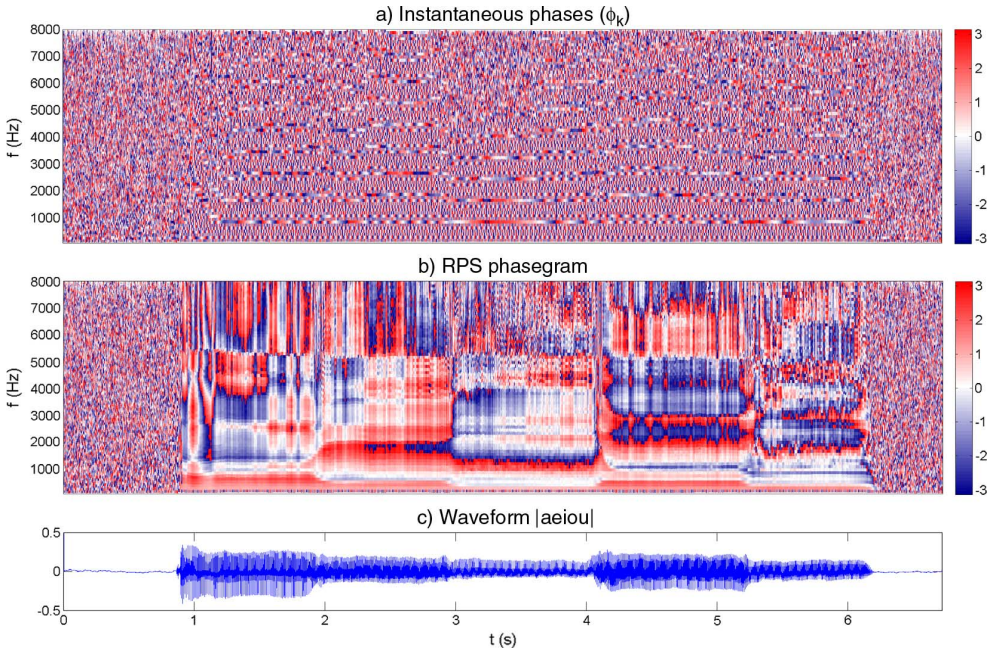


Fig. 3. Phasegrams of a voiced speech segment for five continuous vowels. (a) Instantaneous phases. (b) Relative phase shift. (c) Signal waveform.

every harmonic are then calculated from the instantaneous phase $\Phi_k(t)$ at each analysis instant t_a using

$$\mathrm{RPS}_k = \Phi_k(t_a) - k\Phi_1(t_a). \tag{11}$$

More specifically, this transformation removes the linear phase contribution due to the frequency of every harmonic from the instantaneous phase and allows a clear phase structure to arise, as shown in Fig. 3. The RPS values for voiced segments are illustrated in Fig. 3(b) and show a structured pattern along frequency as the signal evolves.

In order to use RPS values as features for classification and detection of synthetic speech, several important steps must be carried out. These steps were initially developed for an ASR task [45] and are listed below:

1) Due to the variable number of harmonics found in a predefined frequency range, the dimensionality of the vector of RPS values varies from frame to frame. We transform the variable-dimension vectors into fixed-dimension vectors by applying a Mel-scale filter bank with 32 filters.
2) The dimensionality of the RPS vector is very high, if the usual analysis bandwidth is considered. This is problematic for training any statistical model; therefore, RPS values are computed over a frequency range from 0 to 4 kHz and the discrete cosine transform (DCT) is used at the end of the process to decorrelate and reduce the dimensionality.
3) The RPS values in (11) are wrapped phase values and therefore may create discontinuities as shown in Fig. 4(a) and (b). This is also problematic for parameterization. Therefore we unwrap the phase in order to avoid the discontinuities in the RPS envelope.
4) Due to its accumulative, nonlinear nature, the unwrapping process leads to very different RPS envelopes even if they derive from similar initial data as shown in Fig. 4(c) and (d). If we differentiate the unwrapped RPS envelope the accumulative effect is eliminated, the range of the curve is limited to $[-\pi, \pi]$, and thus similarities between envelopes are more properly perceived. This can be seen in Fig. 4(e) and (f).

In order to develop a classifier for synthetic speech, we compute 20 coefficients per speech frame according to steps 1–4. The mean of the differentiated unwrapped RPS (i.e. the mean slope of the unwrapped RPS) has been removed before calculating the DCT and added as a parameter, resulting in a total of 21 coefficients per frame which are used as a feature vector, $\mathbf{y}_t$ for the classifier. Here only voiced segments of the signals have been used, because there is no useful phase information in unvoiced frames. The voiced/unvoiced decision is made using the cepstrum-based pitch detection (CDP) algorithm [47]. The RPS values are then extracted using a 10 ms frame-rate.

For the SSD, we use a 32-component density GMM in the classifier trained on RPS feature vectors extracted from human and synthetic speech signals. Detection of synthetic speech occurs once the speaker verification system has accepted the identity (see Fig. 5)—currently, we see no need to apply the SSD if the SV system has rejected the identity. If an identity claim, C is accepted, we compute the log-likelihood ratio

$$\Lambda_{\mathrm{RPS}}(\mathbf{Y}) = \log p(\mathbf{Y}|\lambda_{C,\mathrm{human}}) - \log p(\mathbf{Y}|\lambda_{C,\mathrm{synth}}) \tag{12}$$

where $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$ is the sequence of RPS feature vectors and $\lambda_{C,\mathrm{human}}$ and $\lambda_{C,\mathrm{synth}}$ represent GMMs of the RPS feature vectors for human and synthetic speech associated with claimant C, respectively. The speaker is then classified as human if $\Lambda_{\mathrm{RPS}}(\mathbf{Y}) > 0$, otherwise it is classified as synthetic.

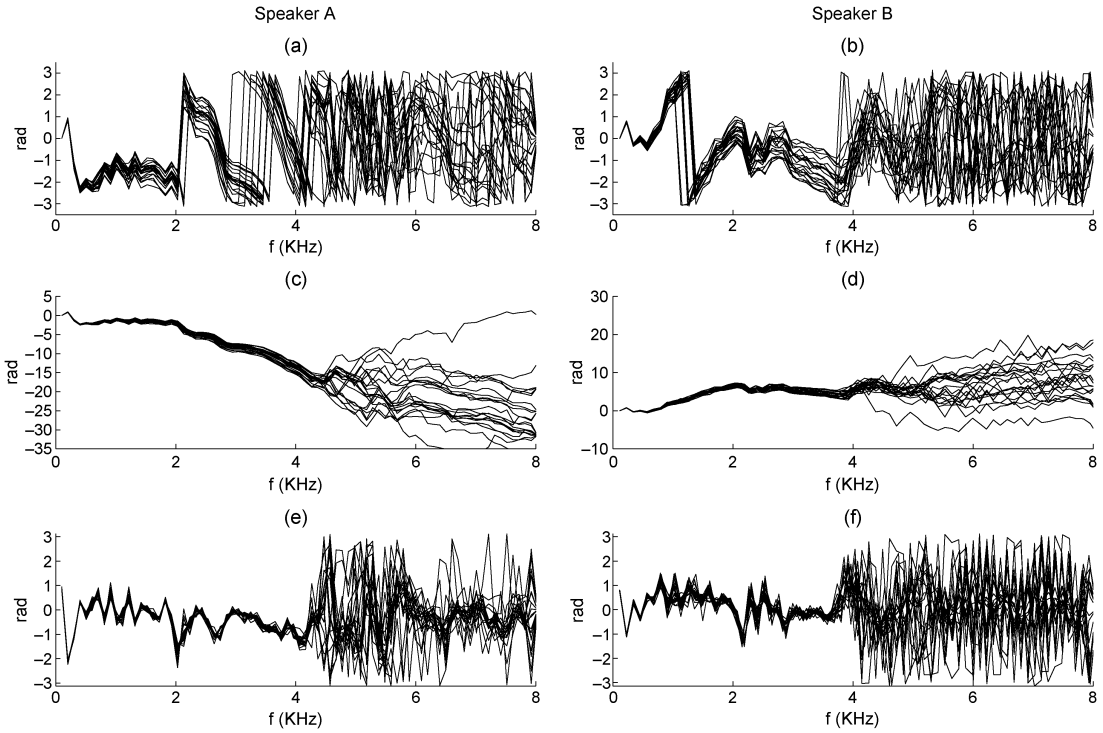


Fig. 4. RPS information for two sustained —i— speech segments of 200 ms (20 frames) by two male speakers: (a)–(b) RPS, (c)–(d) unwrapped RPS, (e)–(f) differentiation of the unwrapped RPS.

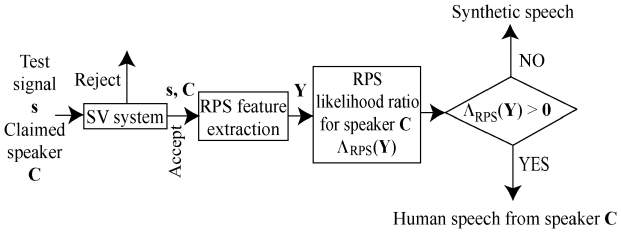


Fig. 5. Proposed system for detection of synthesized speech after speaker verification using phase-based detection.

TABLE I
WALL STREET JOURNAL (WSJ) CORPUS PARTITIONS USED FOR TRAINING AND TESTING OF TEXT-TO-SPEECH (TTS), SPEAKER VERIFICATION (SV), AND SYNTHETIC SPEECH DETECTOR (SSD) SYSTEMS

Human speech (HS)	HS-A train TTS	HS-B train SV train SSD	HS-C test SV test SSD
Synthetic speech (TTS)		TTS-B train SSD	TTS-C test SV test SSD
transCoded speech (CS)		CS-B train SSD	

V. DATA SETS

For this research, we use the WSJ corpus from the Linguistic Data Consortium (LDC) [48]. Although the WSJ corpus is not a standard corpus for SV research, it is one of the few corpora that provides several hundred speakers and sufficiently long signals required for constructing each of the components within the TTS, SV, and SSD systems [49]. From the corpus, we chose the predefined official training data set, SI-284, that includes both WSJ0 and WSJ1 as material data. The SI-284 set has a total of 81 hours of speech data uttered by 283 speakers and was partitioned into three disjoint "human speech" subsets HS-A, HS-B, and HS-C, as shown in Table I. Subset HS-A was used to train the TTS system described in Section III, subset HS-B was used to train the SV and SSD systems described in Sections II and IV-B, and subset HS-C was used to test the SV and SSD systems. Once trained, the TTS system was used to generate the synthetic speech subsets TTS-B and TTS-C as shown in Table I which are used to train the SSD and test the SV and SSD systems, respectively. These different subsets were used to avoid any overlapping of data sets and associated cross-corpus negative effects while attempting to simulate realistic imposture scenarios[2]

Training the SSD with synthetic speech has a practical disadvantage, that is, a TTS synthesizer has to be trained for each speaker in the SV system. Therefore, we have also evaluated a more practical method that uses the STRAIGHT vocoder to transcode the human speech signal as a surrogate for TTS-generated (synthesized) speech. By transcoding, the human speech signal is parametrized using a vocoder and from this parameterization, the speech signal is reconstructed in a process similar to that in the TTS speech generation component. The transcoded human speech signal has artifacts similar to those in the synthetic speech signal which can be useful for simplifying the training of the SSD. In order to evaluate this approach, we transcoded subset HS-B and created the CS-B "coded speech" subset as shown in Table I. By using CS-B instead of TTS-B to

[2]In future work, the average voice model of the TTS should be derived from a different corpus.

TABLE II
ACCEPTANCE RATES FOR HUMAN SPEECH (TRUE CLAIMANT) AND SYNTHETIC SPEECH (MATCHED CLAIM) FOR OVERALL SYSTEM CONSISTING OF SPEAKER VERIFICATION (SV) AND SYNTHETIC SPEECH DETECTOR (SSD). IDEALLY THE SYSTEM HAS 100% ACCEPTANCE RATE FOR HUMAN SPEECH, TRUE CLAIM AND 0% FOR SYNTHETIC SPEECH, MATCHED CLAIM

	GMM-UBM	SVM
Without SSD		
Acceptance rate for human, true claim	99.7%	100%
Acceptance rate for synth, matched claim	85.5%	81.3%
With SSD trained on TTS-B		
Acceptance rate for human, true claim	99.6%	100%
Acceptance rate for synth, matched claim	0.0%	0.0%
With SSD trained on CS-B		
Acceptance rate for human, true claim	99.6%	100%
Acceptance rate for synth, matched claim	8.8%	8.8%
With SSD (set for EER) trained on CS-B		
Acceptance rate for human, true claim	96.8%	97.2%
Acceptance rate for synth, matched claim	2.5%	2.5%

train the SSD, all system components (TTS, SV, SSD) can be trained using only human speech.

Since each speaker included in the SI-284 set has different speech durations, we used varying lengths (73 s to 27 min) of training signals from subset HS-A to construct and adapt the TTS system to each speaker. Some speakers have larger amounts of data than those we can practically collect for the imposture against the SV system.

VI. EXPERIMENTS AND RESULTS

A. Evaluation of Speaker Verification Systems

For the two SV systems, we have trained using ≈90 s speech signals from subset HS-B and tested using ≈30 s signals from subsets HS-C and TTS-C. Training signals for the SVM SV system were segmented into eight utterances per speaker and used to construct GMM supervectors as described in Section II-A. The evaluation for human speech was designed so that each test utterance has an associated true claim and 282 false claims yielding a total of 283^2 tests. The EERs are 0.284%, 0.002% for the GMM-UBM, SVM system respectively. The low EERs (<0.3% for both SV systems) are due to the ideal nature of the recordings in the WSJ corpus and the accuracy of the SV systems. Table II row 2 shows the acceptance rates of the SV systems under human speech for true claims as 99.7%, 100% for the GMM-UBM, SVM system respectively.

The evaluation for synthetic speech was designed so that each test utterance has an associated matched claim yielding 283 tests for imposture. (In a realistic imposture scenario, a speech signal targeted at a specific speaker will be synthesized and a claim only for that speaker will be submitted, i.e. matched claim.) For both SV systems, the decision thresholds are chosen for EER under human speech signal tests. Table II row 3 shows the results where we see over 81% of synthetic speech signals with an associated matched claim will be accepted by the SV systems. As described in an earlier paper, this result is due to significant overlap in the score distributions for human and synthetic speech, as shown in Fig. 6 [24]. Thus, adjustments in decision thresholding or standard score normalization techniques cannot

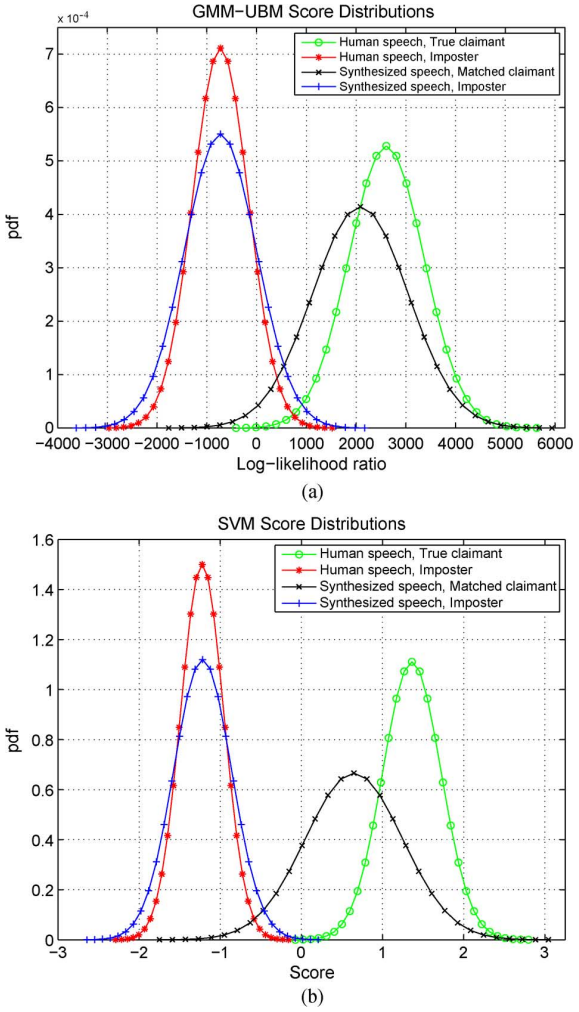


Fig. 6. Approximate score distributions for (a) GMM-UBM and (b) SVM using GMM supervectors SV systems with human and synthesized speech. Distributions for human speech, true claimant (green lines, o) and synthesized speech, matched claimant (black lines, x) have significant overlap leading to a 81% acceptance rate for synthetic speech with matched claims.

differentiate between true and matched claims originating from human and synthesized speech [50], [51]. For completeness in Fig. 6, we show the score distributions for synthesized speech, false claim (imposter) even though in the imposture scenario, only matched claims would be submitted.

B. Evaluation of Synthetic Speech Detector

We trained the SSD, described in Section IV-B, on human speech using HS-B and synthetic speech using TTS-B as in Table I and evaluated classifier accuracy with human speech from HS-C and synthetic speech from TTS-C. These results are shown in Table III row 1 where we find 100% accuracy in classifying a speech signal as either human or synthetic. As mentioned earlier, constructing synthetic voices for each human registered in the SV system is not very practical, so we trained the SSD using transcoded human speech CS-B as a surrogate for synthetic speech. These results are shown in Table III where we find that with the decision threshold set to zero, human speech signals are classified with 100% accuracy and synthetic speech

TABLE III
ACCURACY RATES FOR CLASSIFICATION OF HUMAN AND SYNTHETIC SPEECH. CLASSIFIER IS TRAINED WITH HUMAN SPEECH HS-B AND EITHER TTS-B OR CS-B FOR SYNTHETIC SPEECH. CLASSIFIER IS TESTED USING HS-C AND TTS-C. RESULTS ARE BASED ON A ZERO THRESHOLD FOR LOG-LIKELIHOOD RATIO (12) AND INCLUDE AN ADDITIONAL RESULT FOR CS-B WHERE THRESHOLD IS ADJUSTED FOR EER

Training Data	Human Speech (HS-C)	Synthetic Speech (TTS-C)
HS-B/TTS-B	100%	100%
HS-B/CS-B	100%	90.10%
HS-B/CS-B (EER)	97.17%	97.17%

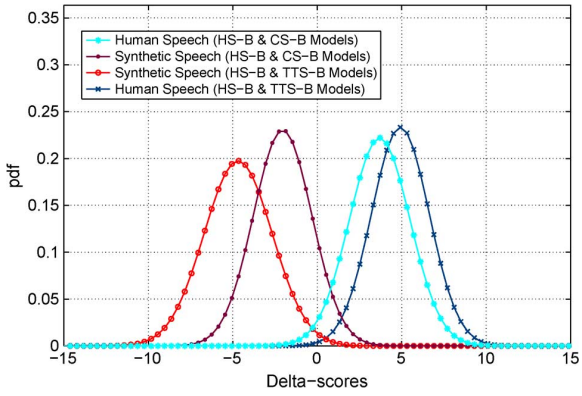


Fig. 7. Approximate distributions for the classifier scores, $\Lambda_{\mathrm{RPS}}(\mathbf{Y})$ when tested with human and synthetic speech. The models for Human Speech are trained with HS-B. Blue and red curves show the classifier performance when the models for synthetic speech are trained using TTS-B. Cyan and magenta curves show the classifier performance when the models for synthetic speech are trained using coded speech CS-B. Both classifiers were tested with human speech HS-C and synthetic speech TTS-C.

signals are classified with 90.10% accuracy. With the decision threshold set at 1.65 for EER, we find 97.17% accuracy in classifying a speech signal as either human or synthetic. Approximate distributions for the classifier scores, $\Lambda_{\mathrm{RPS}}(\mathbf{Y})$ are shown in Fig. 7 where we see that with transcoded speech (CS-B models) it is necessary to adjust the decision threshold slightly upward for EER.

C. Evaluation of Sensitivity of Synthetic Speech Detector

In the evaluation of the SSD in Section VI-B, we have assumed that the same vocoder (STRAIGHT) and phase model (minimum phase) have been used in both training and test stages. Although STRAIGHT is the most popular approach to vocoding and the minimum phase model is normally used, in a real scenario, a different type of vocoder (e.g., [52]) and phase model could be used for imposture. Therefore, we have investigated sensitivity to vocoder mismatch by experimenting with a simple vocoder which uses pulse/white noise excitation and the MLSA filter [42], [53]. We have also investigated sensitivity to phase model mismatch by experimenting with group delay modification [54].

Because the SSD features are entirely phased-based, any mismatch between vocoder and phase model which produces different phase characteristics, may render the classifier's ability to detect synthetic speech unreliable. We have observed this effect in informal tests. When we train the SSD using the aforementioned vocoders, the accuracy of synthetic speech detection falls from 90.1% obtained with the original STRAIGHT vocoder, to 6.3% when training with the pulse/white noise excitation vocoder, and to 50% when training with the group delay modification vocoder. In all cases, the tests were done using TTS-C. On the other hand, classifier accuracy for the human speech remains at 100%. In order to address this issue, future research of a vocoder-independent and phase-adaptive approach such as MAP adaptation of the RPS-GMMs used for the SSD system, will have to be undertaken.

D. Evaluation of Overall System

Next, we evaluated the *overall* system which includes the SV and SSD systems as illustrated in Fig. 5. Using the proposed SSD trained on TTS-B, we see in Table II rows 5–6, there is only a slight 0.1% drop to 99.6% in the acceptance rate for human speech for the GMM-UBM system and no change with the SVM system while the acceptance rate for synthetic speech is now reduced to 0% from over 81% thus clearly illustrating the effectiveness of the SSD using RPS features.

Training the SSD on CS-B, we see in Table II no change in the acceptance rate for human speech compared to training with TTS-B and an acceptance rate for synthetic speech of 8.8% for both SV systems. Finally, adjusting the decision threshold in the SSD for EER, we see in Table II a reduction in acceptance rate for synthetic speech to 2.5% with a slight decrease in acceptance rate for human speech (around 97%). From these results, we conclude that the SSD trained on transcoded speech can drastically reduce the number of accepted matched claims associated with synthetic speech, with only a slight loss in SV accuracy for human speech. Thus the proposed SSD using RPS features is an accurate and effective method for securing the SV systems against imposture using synthetic speech.

E. Evaluation of an Integrated System

Essentially, Fig. 5 represents a system consisting of two separate classifiers: SV using MFCC features and SSD using RPS features. These classifiers can be integrated into a single classifier which uses vectors composed of both MFCC and RPS features. We extracted 53-D feature vectors by concatenating the MFCC feature vector (32-D) described in Section II-A with the RPS feature vector (21-D) described in Section IV-B. In the first simulation, the GMM-UBM and SVM classifiers based on the MFCC-RPS feature vectors were trained using HS-B only and in the second simulation, were trained using both HS-B and TTS-B datasets. When using HS-B and TTS-B, synthetic speakers were treated as imposters in the training stage. The systems were evaluated using HS-C and TTS-C datasets and the results are shown in Table IV.

To begin the evaluation of the integrated system, we first must establish whether the addition of RPS features compromises SV accuracy when the system is trained and tested only with human speech. For this case, the addition of the RPS features slightly raises the EER to 0.35%, 0.02% for the GMM-UBM, SVM system, respectively, as compared to the SV system which uses MFCC features only. The acceptance rates for true claims (99.7% for GMM-UBM, 100.0% for SVM) remain the same as

TABLE IV
ACCEPTANCE RATES FOR HUMAN SPEECH (TRUE CLAIMANT) AND SYNTHETIC SPEECH (MATCHED CLAIM) FOR THE INTEGRATED SYSTEM (SINGLE CLASSIFIER) WHICH USES VECTORS COMPOSED OF BOTH MFCC AND RPS FEATURES. IDEALLY THE SYSTEM HAS 100% ACCEPTANCE RATE FOR HUMAN SPEECH, TRUE CLAIM AND 0% FOR SYNTHETIC SPEECH, MATCHED CLAIM

	GMM-UBM	SVM
Integrated System Trained on HS-B		
Acceptance rate for human, true claim	99.7%	100%
Acceptance rate for synthetic, matched claim	88.7%	56.2%
Integrated System Trained on HS-B and TTS-B		
Acceptance rate for human, true claim	99.3%	100%
Acceptance rate for synthetic, matched claim	40.6%	3.5%

compared to the SV system which uses MFCC features only. These results thus demonstrate that the addition of the RPS features does not appreciably change SV accuracy under human speech.

Earlier, we illustrated the imposture problem by demonstrating that when the SV systems using MFCC features were trained on human speech and tested with synthetic speech, the acceptance rates for matched claims were high (85.5%, 81.3% for the GMM-UBM, SVM systems, respectively). With the integrated system (GMM-UBM classifier), the acceptance rate for matched claims increases to 88.7% from 85.5%. On the other hand, the SVM system shows a notable drop in the acceptance rate to 56.2% from 81.3%. Unfortunately, both acceptance rates for synthetic speech with matched claims are still unacceptably high.

Next, we compare the integrated system trained with human and synthetic speech to the system composed of separate SV and SSD stages in Fig. 5. When the integrated system is tested with human speech, the acceptance rates for true claims drops slightly to 99.3% for the GMM-UBM system and remains the same 100% for the SVM system. When the GMM-UBM integrated system is evaluated with synthetic speech, the acceptance rate for matched claims is 40.6%. Not surprisingly, the GMM-UBM integrated system appears to have an average performance with synthetic speech between the stand-alone rates of the SV using MFCCs (85.5%) and the SSD using RPS (0.0%). When the SVM integrated system is evaluated with synthetic speech, the acceptance rate for matched claims is 3.5% which is still higher than for the system composed of separate SV and SSD stages which is also 0.0% (Table II, row 6). For both GMM-UBM and SVM integrated systems, inclusion of synthetic speech signals in training lowers the acceptance rates for synthetic speech, matched claims by around 50% (from 88.7% to 40.6% for GMM-UBM and from 56.2% to 3.5% for SVM). However, these results demonstrate that the proposed system composed of separate SV and SSD classifiers (Fig. 5) performs better than the integrated system. Nevertheless, the performance of the integrated SVM system is notable in that it does not use a separate synthetic impostor model for each speaker as the separate SSD does. Since training with CS-B leads to a less accurate model for synthetic speech than with TTS-B (see Table II, rows 6, 9, and 12) and results for the integrated system trained with TTS-B are worse than with the separate system, the integrated system is not trained with CS-B and evaluated.

VII. CONCLUSION

In this paper, we have evaluated the vulnerability of SV to imposture using synthetic speech. Using the WSJ corpus and two different SV systems (GMM-UBM and SVM using GMM supervectors), we have shown that with state-of-the-art speech synthesis, over 81% of matched claims, i.e. a synthetic speech signal matched to a targeted speaker and an identity claim of that same speaker, are accepted. Thus, despite the excellent performance of the SV systems under human speech, the quality of synthesized speech is high enough to allow these synthesized voices to pass for true human claimants and hence poses a potential security problem.

We have proposed a novel SSD based on RPS features. Although the SSD can detect human and synthetic speech with 100% accuracy, training requires that a TTS synthesizer be constructed for each speaker in the SV system which is not practical. Therefore, we have proposed using transcoded speech as a surrogate for synthetic speech in training the SSD. Our results show that we can reduce the acceptance rate of synthetic speech, matched claims from over 81% to 2.5%, with a less than 3% drop in the acceptance rate for human speech, true claimants. However, the system is sensitive to the vocoder used: the same vocoder used by the impostor must be used to train the system. The investigation of vocoder-independent techniques is left for future work.

REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[2] T. Matsui and S. Furui, "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model," *Speech Commun.*, vol. 17, no. 1–2, pp. 109–116, Aug. 1995.

[3] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proc. Int. Symp. Intell. Multimedia, Video, Speech Process.*, Oct. 2004, pp. 145–148.

[4] K. Sullivan and J. Pelecanos, "Revisiting carl bildt's impostor: Would a speaker verification system foil him?," in *Audio- and Video-Based Biometric Person Authentication*, ser. Lecture Notes Computer Science, J. Bigun and F. Smeraldi, Eds. Berlin/Heidelberg, Germany: Springer, 2001, vol. 2091, pp. 144–149.

[5] D. Genoud and G. Chollet, "Speech pre-processing against intentional imposture speaker recognition," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, Dec. 1998, vol. 2, pp. 105–108.

[6] B. L. Pellom and J. H. L. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1999, pp. 837–840.

[7] J. Lindberg and M. Blomberg, "Vulnerability speaker verification-a study of possible technical imposter techniques.," in *Proc. Eur. Conf. Speech Commun. Technol. (Eurospeech)*, 1999, vol. 3, pp. 1211–1214.

[8] D. Matrouf, J.-F. Bonastre, and C. Fredouille, "Effect of speech transformation on impostor acceptance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2006, vol. 2, pp. 933–936.

[9] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Proc. Int. Speech Commun. Assoc. (Interspeech)*, Apr. 2007, pp. 2053–2056.

[10] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Transfer function-based voice transformation for speaker recognition," in *Proc. IEEE Speaker Lang. Recogn. Workshop (Odyssey)*, Jun. 2006, pp. 1–6.

[11] M. Farrus, D. Erro, and J. Hern, "Speaker recognition robustness to voice conversion," *IV J. Reconocimiento Biometrico de Personas*, pp. 73–82, Sep. 2008.

[12] L.-J. Bo, "Forensic voice identification France," *Speech Commun.*, vol. 31, no. 2–3, pp. 205–224, 2000.

[13] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of HMM-based speaker verification systems against imposture using synthetic speech," in *Proc. Eur. Conf. Speech Commun. Technol. (Eurospeech)*, 1999, pp. 1223–1226.

[14] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 1996, pp. 389–392.

[15] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1997, pp. 1611–1614.

[16] T. Masuko, K. Tokuda, and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, 2000, vol. 2, pp. 302–305.

[17] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleaua, S. Meignier, T. Merlin, J. Ortega-Garcia, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Process.*, vol. 4, pp. 430–451, 2004.

[18] T. Kinnunen, E. Karpov, and P. Franti, "Real-time speaker identification and verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 1, pp. 277–288, Jan. 2006.

[19] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.

[20] C. Longworth and M. Gales, "Combining derivative and parametric kernels for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 748–757, May 2009.

[21] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[22] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Speech, Audio, Lang. Process.*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[23] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis-Analysis and application of TTS systems built on various ASR corpora," *IEEE Trans. Speech, Audio, Lang. Process.*, vol. 18, no. 5, pp. 984–1004, Jul. 2010.

[24] P. L. De Leon, V. R. Apsingekar, M. Pucher, and J. Yamagishi, "Revisiting the security of speaker verification systems against imposture using synthetic speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2010, pp. 1798–1801.

[25] P. L. De Leon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Proc. IEEE Speaker Lang. Recogn. Workshop (Odyssey)*, 2010, pp. 151–158.

[26] P. L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, 2011, pp. 4844–4847.

[27] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," in *Proc. Eur. Conf. Speech Commun. Technol. (Eurospeech)*, 2001, pp. 759–762.

[28] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. IEEE Speaker Lang. Recogn. Workshop (Odyssey)*, 2001, pp. 213–218.

[29] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using GMM supervector kernel and NAP variability compensation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2006, pp. 97–100.

[30] A. Black and N. Cambpbell, "Optimising selection of units from speech database for concatenative synthesis," *Proc. Eurospeech-95*, pp. 581–584, Sep. 1995.

[31] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard challenge 2008," in *Proc. Blizzard Challenge*, Sep. 2008 [Online]. Available: http://festvox.org/blizzard/bc2008/summary_Blizzard2008.pdf

[32] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda, "The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system the 2008 Blizzard Challenge," in *Proc. Blizzard Challenge*, Sep. 2008 [Online]. Available: http://festvox.org/blizzard/bc2008/hts_Blizzard2008.pdf

[33] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure sounds," *Speech Commun.*, vol. 27, pp. 187–207, 1999.

[34] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

[35] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.

[36] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Japan (E)*, vol. 21, pp. 79–86, Mar. 2000.

[37] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, Oct. 1996, pp. 1137–1140.

[38] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75–98, 1998.

[39] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Speech, Audio, Lang. Process.*, vol. 17, no. 1, pp. 66–83, 2009.

[40] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[41] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5–6, pp. 453–468, 1990.

[42] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for Mel-cepstral analysis of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 1992, pp. 137–140.

[43] T. F. Quatieri, *Discrete-Time Speech Signal Processing Principles and Practice*. Upper Saddle River, NJ: Prentice-Hall, 2002.

[44] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez, "Simple representation of signal phase for harmonic speech models," *Electron. Lett.*, vol. 45, pp. 381–383, 2009.

[45] I. Saratxaga, I. Hernaez, I. Odriozola, E. Navas, I. Luengo, and D. Erro, "Using harmonic phase information to improve ASR rate," in *Proc. Int. Speech Commun. Assoc. (Interspeech)*, 2010, pp. 1185–1188.

[46] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. dissertation, Ecole Nationale Supérieure des Télécommunications, Paris, France, 1996.

[47] I. Luengo, I. Saratxaga, E. Navas, I. Hernaez, J. Sanchez, and I. Sainz, "Evaluation of pitch detection algorithms under real conditions," in *Proc. ICASSP '07*, Honolulu, HI, Apr. 2007, pp. 1057–1060.

[48] Wall Street Journal Corpus, 2010. [Online]. Available: http://www.ldc.upenn.edu

[49] D. B. Paul and J. M. {Baker}, "The design for the Wall Street Journal-based CSR corpus," in *Proc. Workshop Speech Natural Lang.*, 1992, pp. 357–362.

[50] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," in *Proc. IEEE. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 1988, vol. 1, pp. 595–598.

[51] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for test-independent speaker verification system}," *Digital Signal Process.*, vol. 10, no. 1, pp. 42–54, 2000.

[52] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 153–165, Jan. 2011.

[53] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electron. Commun. Japan (Part I: Commun.)* vol. 66, no. 2, pp. 10–18, 1983 [Online]. Available: http://dx.doi.org/10.1002/ecja.4400660203

[54] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *Proc. Models Anal. Vocal Emissions for Biomed. Applicat. (MAVEBA)*, 2001, pp. 1–6.

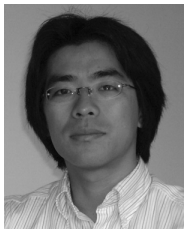
Phillip L. De Leon (SM'03) received the B.S. degree in electrical engineering and the B.A. degree in mathematics from the University of Texas at Austin in 1989 and 1990, respectively, and the M.S. and Ph.D. degrees in electrical engineering from the University of Colorado at Boulder in 1992 and 1995, respectively.

In 2002, he was a Visiting Professor in the Department of Computer Science, University College Cork, Cork, Ireland. In 2008, he was selected by the U.S. State Department as a Fulbright Faculty Scholar and served as a Visiting Professor at the Technical University in Vienna (TU-Vienna). Currently, he is a Professor and Associate Department Head in the Klipsch School of Electrical and Computer Engineering and Director of the Advanced Speech and Audio Processing Laboratory at New Mexico State University, Las Cruces. His research interests are in speech-signal processing, embedded systems, and pattern recognition and machine learning.

Michael Pucher (M'12) received the Ph.D. degree from the Graz University of Technology, Graz, Austria, in 2007 with a thesis on semantic language modeling for speech recognition.

He is a Senior Researcher and Project Manager at the Telecommunications Research Center Vienna (FTW), Vienna, Austria. His research interests are speech synthesis and recognition, multimodal dialog systems, and sensor fusion. He has authored and coauthored more than 30 refereed papers in international conferences and journals. In 2010 he was involved in the commercial development of Leopold, the first synthetic voice for Austrian German.

Dr. Pucher was awarded a research grant from the Austrian Science Fund (FWF) for the project "Adaptive Audio-Visual Dialect Speech Synthesis" (AVDS) in 2011. A list of publications and a detailed CV can be found on http://userver.ftw.at/~pucher.



Junichi Yamagishi received the Ph.D. degree from the Tokyo Institute of Technology, Tokyo, Japan, in 2006 with a thesis which pioneered the use of adaptation techniques in HMM-based speech synthesis.

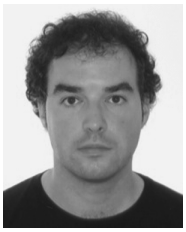
Since 2006, he has been with the Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, U.K. In addition to authoring and coauthoring over 80 refereed papers in international journals and conferences, his work has led directly to two large-scale EC FP7 projects and two collaborations based around clinical applications of this technology. He is an external member of the Euan MacDonald Centre for MND Research in Edinburgh and a Visiting Associate Professor of Nagoya Institute of Technology, Japan.

Dr. Yamagishi was awarded the Tejima Prize as the best Ph.D. thesis of the Tokyo Institute of Technology in 2007. In 2010, he was awarded the Itakura Prize (Innovative Young Researchers Prize) from the Acoustical Society of Japan for his achievements in adaptive speech synthesis.

Inma Hernáez received the telecommunications engineering degree from the Universitat Politecnica de Catalunya, Barcelona, Spain, and the Ph.D. degree in telecommunications engineering from the University of the Basque Country, Bilbao, Spain, in 1987 and 1995, respectively.

She is a Full Professor in the Electronics and Telecommunication Department, Faculty of Engineering, University of the Basque Country, Bilbao, Spain, in the area of signal theory and communications and founding member and director of the Aholab Signal Processing Laboratory. Her research interests include signal processing and all aspects related to speech processing. She is also interested in the development of speech resources and technologies for the Basque language.

Dr. Hernáez is a member of the International Speech Communication Association (ISCA), the Spanish thematic network on Speech Technologies (RTTH), and the European Center of Excellence on Speech Synthesis (http://www.ecess.eu).



Ibon Saratxaga received the telecommunications engineering degree from the University of the Basque Country, Bilbao, Spain, in 1995.

Since 2005, he has been a Researcher at the Aholab Signal Processing Laboratory, University of the Basque Country, Bilbao, Spain. He is currently teaching at the Faculty of Engineering in Bilbao. He has participated as junior research in several funded research projects. His research interests include speech coding and synthesis, with a focus on the study of the harmonic phase of the speech signal.

Mr. Saratxaga is a member of the International Speech Communication Association (ISCA), the Spanish thematic network on Speech Technologies (RTTH), and the European Center of Excellence of Speech Synthesis (http://www.ecess.eu).

Principal references

De Leon, P. L., Pucher, M., Yamagishi, J., Hernaez, I., and Saratxaga, I. (2012). Evaluation of speaker verification security and detection of HMM-based synthetic speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 20:2280–2290.

Hollenstein, J., Pucher, M., and Schabus, D. (2013). Visual control of hidden-semi-Markov-model based acoustic speech synthesis. In *International Conference on Auditory-Visual Speech Processing (AVSP 2013)*, pages 31–35, Annency, France.

Pucher, M., Schabus, D., and Yamagishi, J. (2010a). Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and sighted listeners. In *INTERSPEECH 2010*, pages 2186–2189, Makuhari, Japan.

Pucher, M., Schabus, D., Yamagishi, J., Neubarth, F., and Strom, V. (2010b). Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis. *Speech Communication*, 52:164–179.

Pucher, M., Toman, M., Schabus, D., Valentini-Botinhao, C., Yamagishi, J., Zillinger, B., and Schmid, E. (2015). Influence of speaker familiarity on blind and visually impaired children's perception of synthetic voices in audio games. In *INTERSPEECH 2015*, pages 1625–1629, Dresden, Germany.

Schabus, D., Pucher, M., and Hofer, G. (2012). Speaker-adaptive visual speech synthesis in the HMM-framework. In *INTERSPEECH 2012*, pages 979–982, Portland, USA.

Schabus, D., Pucher, M., and Hofer, G. (2014). Joint audiovisual hidden semi-Markov model-based speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):336–347.

Toman, M., Pucher, M., Moosmüller, S., and Schabus, D. (2015). Unsupervised and phonologically controlled interpolation of Austrian German language varieties for speech synthesis. *Speech Communication*, 72:176–193.

Valentini-Botinhao, C., Toman, M., Pucher, M., Schabus, D., and Yamagishi, J. (2014). Intelligibility analysis of fast synthesized speech. In *INTERSPEECH 2014*, pages 2922–2926, Singapore.

Valentini-Botinhao, C., Toman, M., Pucher, M., Schabus, D., and Yamagishi, J. (2015). Intelligibility of time-compressed synthetic speech: Compression method and speaking style. *Speech Communication*, 74:52–64.

Secondary references

Anegg, H., Dangl, T., Jank, M., Niklfeld, G., Pucher, M., Schatz, R., Simon, R., and Wegscheider, F. (2004). Multimodal interfaces in mobile devices - the MONA project. In *Proceedings of the Workshop on Emerging Applications for Wireless and Mobile Access. 13th International World Wide Web Conference (WWW 2004)*, New York, USA.

Baillie, L., Pucher, M., and Képesi, M. (2004). A supportive multimodal mobile robot for the home. In *User-centered interaction paradigms for universal access in the information society*, pages 375–383. Springer.

Bruss, M. (2008). Quantitative und phonetische Analyse von nicht-linguistischen Partikeln in spontan gesprochener Sprache der Wiener Soziolekte. Master's thesis, Universität des Saarlandes, Saarbrücken, Germany.

Cereproc (2010). Leopold at webshop of Cereproc. `https://www.cereproc.com/de/slandingde`.

De Leon, P. L., Apsingekar, V. R., Pucher, M., and Yamagishi, J. (2010a). Revisiting the security of speaker verification systems against imposture using synthetic speech. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 1798–1801.

De Leon, P. L., Hernaez, I., Saratxaga, I., Pucher, M., and Yamagishi, J. (2011). Detection of synthetic speech for the problem of imposture. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 4844–4847, Dallas, USA.

De Leon, P. L., Pucher, M., and Yamagishi, J. (2010b). Evaluation of the vulnerability of speaker verification to synthetic speech. In *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey)*, pages 151–158.

Esposito, A. (2006). Cross-modal analysis of verbal and non-verbal communication. `http://www.cost.eu/COST_Actions/ict/2102`.

EUCOG (2015). EUCOG III: European network for the advancement of artificial cognitive systems, interaction and robotics. `http://www.eucognition.org/`.

Fröhlich, P. and Pucher, M. (2005). Combining Speech and Sound in the User Interface. In *Proc. Workshop of ICAD 2005, Limerick*.

FTW (2004). MONA - Mobile multimodal next generation applications.

FTW (2006). TIDE - Testbed for interactive dialog system evaluation.

Hollenstein, J. (2013). Visual control of audio-visual speech synthesis. Master's thesis, Vienna University of Technology, Austria.

Kranzler, C. (2008). Text-to-speech engine with Austrian German corpus. Master's thesis, Graz University of Technology, Austria.

Kranzler, C., Pernkopf, F., Muhr, R., Pucher, M., and Neubarth, F. (2009). Text-to-speech engine with Austrian German corpus. In *Proceedings of the XIII International conference Speech and Computer (SPECOM 2009)*, pages 2341–2344, St. Petersburg, Russia.

Möller, S., Engelbrecht, K.-P., Pucher, M., Fröhlich, P., Huo, L., Heute, U., and Oberle, F. (2007). TIDE: A testbed for interactive spoken dialogue system evaluation. In *Proc. Intl. Conf. Speech and Computers*, volume 6.

Möller, S., Engelbrecht, K.-P., Pucher, M., Fröhlich, P., Huo, L., Heute, U., and Oberle, F. (2008). A new testbed for semi-automatic usability evaluation and optimization of spoken dialogue systems. In *Usability of Speech Dialog Systems*, pages 81–103. Springer.

Neubarth, F., Pucher, M., and Kranzler, C. (2008). Modeling Austrian dialect varieties for TTS. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008)*, pages 1877–1880, Brisbane, Australia.

Niklfeld, G., Anegg, H., Pucher, M., Schatz, R., Simon, R., Wegscheider, F., Gassner, A., Jank, M., and Pospischil, G. (2005a). Device independent mobile multimodal user interfaces with the MONA multimodal presentation server. In *Proceedings of the Eurescom summit 2005 on Ubiquitous Services and Applications*, Heidelberg, Germany.

Niklfeld, G., Finan, R., and Pucher, M. (2001a). Architecture for adaptive multimodal dialog systems based on VoiceXML. In *INTERSPEECH 2001*, pages 2341–2344.

Niklfeld, G., Finan, R., and Pucher, M. (2001b). Component-based multimodal dialog interfaces for mobile knowledge creation. In *Proceedings of the workshop on Human Language Technology and Knowledge Management-Volume 2001*, page 16. Association for Computational Linguistics.

Niklfeld, G., Finan, R., and Pucher, M. (2001c). Multimodal interface architecture for mobile data services. In *Proceedings of TCMC2001 Workshop on Wearable Computing, Graz, Austria.*

Niklfeld, G., Pucher, M., Finan, R., and Eckhart, W. (2002a). Kombinierte Sprache/Display-Schnittstellen für mobile Datendienste. *Praxis der Informationsverarbeitung und Kommunikation*, 25(4):196–201.

Niklfeld, G., Pucher, M., Finan, R., and Eckhart, W. (2002b). Mobile multi-modal data services for GPRS phones and beyond. In *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, pages 337–342. IEEE.

Niklfeld, G., Pucher, M., Finan, R., and Eckhart, W. (2002c). Steps towards multimodal data services in GPRS and in UMTS or WLAN networks. In *ISCA Tutorial and Research Workshop (ITRW) on Multi-Modal Dialogue in Mobile Environments*.

Niklfeld, G., Pucher, M., Finan, R., and Eckhart, W. (2005b). A path to multimodal data services for telecommunications. In Minker, W., Bühler, D., and Dybkjr, L., editors, *Spoken Multimodal Human-Computer Dialogue in Mobile Environments*, volume 28 of *Text, Speech and Language Technology*, pages 149–167. Springer Netherlands.

Pucher, M. (2001). Formale Wahrheitstheorien nach Alfred Tarski. Master's thesis, University of Vienna, Austria.

Pucher, M. (2005). Performance evaluation of WordNet-based semantic relatedness measures for word prediction in conversational speech. In *Proceedings of 6th International Workshop on Computational Semantics (IWCS-6)*, pages 332–342, Tilburg, Netherlands.

Pucher, M. (2007a). *Semantic similarity in automatic speech recogniton for meetings*. PhD thesis, Graz University of Technology, Austria.

Pucher, M. (2007b). Viennese sociolect and dialect synthesis (VSDS). `http://web.archive.org/web/20150827100105/https://portal.ftw.at/projects/vsds`.

Pucher, M. (2007c). Wordnet-based semantic relatedness measures in automatic speech recognition for meetings. In *Association for Computational Linguistics (ACL)*, Prague, Czech Republic.

Pucher, M. (2011). Adaptive audio-visual dialect speech synthesis (AVDS) - FWF P22890-N23. `http://web.archive.org/web/20150827095922/https://portal.ftw.at/projects/avds`.

Pucher, M. (2012a). Acoustic modeling and transformation of varieties for speech synthesis (AMTV) - FWF P23821-N23. `http://web.archive.org/web/20150827100349/https://portal.ftw.at/projects/amtv`.

Pucher, M. (2012b). FAA - The 3rd international symposium on facial analysis and animation. `http://speech.kfs.oeaw.ac.at/faa2012/`.

Pucher, M. (2013). Speech synthesis of auditory lecture books for blind children (SALB) - BMWF Sparkling Science. `http://web.archive.org/web/20150827095313/https://portal.ftw.at/projects/salb`.

Pucher, M. (2015a). A Hidden-Markov-Model (HMM) based Opera Singing Synthesis System for German. Master's thesis, Computer Science, Vienna University of Technology, Austria.

Pucher, M. (2015b). Cognitive User Interfaces. `https://tiss.tuwien.ac.at/course/courseDetails.xhtml?locale=en&windowId=d5b&courseNr=185A08&semester=2015S`.

Pucher, M. (2015c). Computational semantics. `https://tiss.tuwien.ac.at/course/courseDetails.xhtml?windowId=cd2&semester=2014W&courseNr=185A67`.

Pucher, M. (2015d). FAAVSP - The 1st joint conference on facial analysis, animation and auditory-visual speech processing. `http://speech.kfs.oeaw.ac.at/faavsp2015/`.

Pucher, M., Cosker, D., Hofer, G., Berger, M., and Smith, W., editors (2012a). *Proceedings of Facial Analysis and Animation 2012*, number ISBN 978-1-4503-1793-1 in The ACM International Conference Proceedings Series, Vienna, Austria. ACM.

Pucher, M., Cosker, D., Krumhuber, E., and Smith, W., editors (2015a). *Proceedings of FAAVSP - The 1st Joint Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing (FAA part)*, number ISBN 978-1-4503-3530-0 in The ACM International Conference Proceedings Series, Vienna, Austria. ACM.

Pucher, M., Cosker, D., Krumhuber, E., Smith, W., Ouni, S., and Davis, C., editors (2015b). *Proceedings of FAAVSP - The 1st Joint Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing (AVSP part)*, ISCA Online Proceedings, Vienna, Austria.

Pucher, M. and Fröhlich, P. (2005). A user study on the influence of mobile device class, synthesis method, data rate and lexicon on speech synthesis quality. In *Proceedings of the 9th European Conference on Speech Communication and Technology (EUROSPEECH 2005)*, pages 2501–2504, Lisboa, Portugal.

Pucher, M. and Huang, Y. (2005). Latent semantic analysis based language models for meetings. In *MLMI05, 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Edinburgh, UK.

Pucher, M., Huang, Y., and Çetin, Ö. (2006a). Combination of latent semantic analysis based language models for meeting recognition. In *Computational Intelligence 2006, Special Session on "Natural Language Processing for Real Life Applications"*, pages 465–469, San Francisco, USA.

Pucher, M., Huang, Y., and Çetin, Ö. (2006b). Optimization of latent semantic analysis based language model interpolation for meeting recognition. In *5th Slovenian and 1st International Language Technologies Conference*, pages 74–78, Ljubljana, Slovenia.

Pucher, M. and Képesi, M. (2003). Multimodal mobile robot control using speech application language tags. In *Ambient Intelligence*, pages 56–64. Springer.

Pucher, M., Kerschhofer-Puhalo, N., and Schabus, D. (2011). Phone set selection for HMM-based dialect speech synthesis. In *EMNLP 2011: Conference on Empirical Methods in Natural Language Processing. 1st Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties (DIALECTS 2011)*, pages 65–69, Edinburgh, UK.

Pucher, M., Kerschhofer-Puhalo, N., Schabus, D., Moosmüller, S., and Hofer, G. (2012b). Language resources for the adaptive speech synthesis of dialects. In *Proc. of the 7th Congress of the International Society for Dialectology and Geolinguistics. Vienna, Austria*, pages 174–175.

Pucher, M., Neubarth, F., Rank, E., Niklfeld, G., and Guan, Q. (2003a). Combining non-uniform unit selection with diphone based synthesis. In *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, pages 1329–1332, Geneva, Switzerland.

Pucher, M., Neubarth, F., and Schabus, D. (2010a). Design and development of spoken dialog systems incorporating speech synthesis of Viennese varieties. In *Computers Helping People with Special Needs*, pages 361–366. Springer.

Pucher, M., Neubarth, F., and Strom, V. (2010b). Optimizing phonetic encoding for Viennese dialect unit selection speech synthesis. In *COST 2102 conference 2009, LNCS 5967*, pages 207–216, Dublin, Ireland.

Pucher, M., Neubarth, F., Strom, V., Moosmüller, S., Hofer, G., Kranzler, C., Schuchmann, G., and Schabus, D. (2010c). Resources for speech synthesis of Viennese varieties. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 105–108, Valletta, Malta.

Pucher, M. and Schabus, D. (2015). Visio-articulatory to acoustic conversion of speech. In *The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (FAAVSP 2015)*, page Article No. 6, Vienna, Austria.

Pucher, M., Schabus, D., and Hofer, G. (2012c). From Viennese to Austrian German and back again - An algorithm for the realization of a variety-slider. In *Proc. of the 7th Congress of the International Society for Dialectology and Geolinguistics. Vienna, Austria*, pages 176–177.

Pucher, M., Schabus, D., Hofer, G., Kerschhofer-Puhalo, N., and Moosmüller, S. (2012d). Regionalizing virtual avatars - towards adaptive audio-visual dialect speech synthesis. In *CogSys 2012, 5th International Conference on Cognitive Systems*, page 95, Vienna, Austria.

Pucher, M., Schabus, D., Schallauer, P., Lypetskyy, Y., Graf, F., Rainer, H., Stadtschnitzer, M., Sternig, S., Birchbauer, J., Schneider, W., et al. (2010d). Multimodal highway monitoring for robust incident detection. In *13th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 837–842. IEEE.

Pucher, M., Schuchmann, G., and Fröhlich, P. (2008). Regionalized text-to-speech systems: Persona design and application scenarios. In *COST 2102 school*, pages 216–222, Vietri, Italy.

Pucher, M., Tertyshnaya, J., and Wegscheider, F. (2003b). Personal voice call assistant: VoiceXML and SIP in a distributed environment. In *Proceedings of the Workshop on Emerging Applications for Wireless and Mobile Access. 12th International World Wide Web Conference (WWW 2003)*, Budapest, Hungary.

Pucher, M., Türk, A., Ajmera, J., and Fecher, N. (2007). Phonetic distance measures for speech recognition vocabulary and grammar optimization. In *3rd Congress of the Alps Adria Acoustics Association*, pages 2–5.

Pucher, M., Villavicencio, F., and Yamagishi, J. (2016a). Development and evaluation of a statistical parametric synthesis system for operatic singing in German. In *Speech Synthesis Workshop (SSW9)*, page (submitted), Sunnyvale, California.

Pucher, M., Xhafa, V., Dika, A., and Toman, M. (2015c). Adaptive speech synthesis of Albanian dialects. In *Text, Speech, and Dialogue (TSD) 2015*, pages 158–164, Pilsen, Czech Republic.

Pucher, M. and Yamagishi, J. (2014). Acoustic modeling and statistical analysis of Vienna opera singers (opera) - NII internal project. `http://www.sociolectix.org/papers/FTW_PA_Computer_singen_Opern.pdf`.

Pucher, M., Zillinger, B., Toman, M., Schabus, D., Valentini-Botinhao, C., Yamagishi, J., Schmid, E., and Woltron, T. (2016b). Influence of speaker familiarity on blind and visually impaired children's perception of synthetic voices. *Computer Speech and Languge Special Issue on Language and Interaction Technologies for Children*, (submitted).

Saratxaga, I., Hernaez, I., Pucher, M., Navas, E., and Sainz, I. (2012). Perceptual importance of the phase related information in speech. In *INTERSPEECH 2012*, pages 1448–1451, Portland, USA.

Schabus, D. (2009). Interpolation of Austrian German and Viennese dialect / sociolect in HMM-based speech synthesis. Master's thesis, Vienna University of Technology, Austria.

Schabus, D. (2014). *Audiovisual speech synthesis based on hidden Markov models*. PhD thesis, Graz University of Technology, Austria.

Schabus, D. and Pucher, M. (2014a). Bad Goisern and Innervillgraten audio-visual dialect speech corpus (GIDS). `http://speech.kfs.oeaw.ac.at/gids/`.

Schabus, D. and Pucher, M. (2014b). Multi-modal annotated synchronous corpus of speech (MMASCS). `http://speech.kfs.oeaw.ac.at/mmascs/`.

Schabus, D. and Pucher, M. (2015). Comparison of dialect models and phone mappings in HSMM-based visual dialect speech synthesis. In *The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing (FAAVSP 2015)*, pages 84–87, Vienna, Austria.

Schabus, D., Pucher, M., and Hofer, G. (2011). Simultaneous speech and animation synthesis. In *ACM SIGGRAPH 2011 Posters*, Vancouver, BC, Canada.

Schabus, D., Pucher, M., and Hofer, G. (2012). Building a synchronous corpus of acoustic and 3D facial marker data for adaptive audio-visual speech synthesis. In *Proc. LREC*, pages 3313–3316, Istanbul, Turkey.

Schabus, D., Pucher, M., and Hofer, G. (2013). Objective and subjective feature evaluation for speaker-adaptive visual speech synthesis. In *Proceedings of the 12th International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 37–42, Annecy, France.

Schabus, D., Pucher, M., and Hoole, P. (2014). The MMASCS multi-modal annotated synchronous corpus of audio, video, facial motion and tongue motion data of normal, fast and slow speech. In *LREC*, pages 3411–3416.

Toman, M. (2013). SALB - frontend for speech synthesis using HTS voice models. `http://m-toman.github.io/SALB/`.

Toman, M. (2016). *Transformation and interpolation of language varieties in speech synthesis*. PhD thesis, Vienna University of Technology, Austria.

Toman, M. and Pucher, M. (2013). Structural KLD for cross-variety speaker adaptation in HMM-based speech synthesis. In *SPPRA*, Innsbruck, Austria.

Toman, M. and Pucher, M. (2015a). Evaluation of state mapping based foreign accent conversion. In *INTERSPEECH 2015*, pages 304–308, Dresden, Germany.

Toman, M. and Pucher, M. (2015b). An open source speech synthesis frontend for HTS. In *Text, Speech, and Dialogue (TSD) 2015*, pages 291–298, Pilsen, Czech Republic.

Toman, M., Pucher, M., and Schabus, D. (2013a). Austrian German voices for the Festival speech synthesis system. `http://sourceforge.net/projects/at-festival/`.

Toman, M., Pucher, M., and Schabus, D. (2013b). Cross-variety speaker transformation in HSMM-based speech synthesis. In *8th ISCA Speech Synthesis Workshop*, pages 77–81, Barcelona, Spain.

Toman, M., Pucher, M., and Schabus, D. (2013c). Multi-variety adaptive acoustic modeling in HSMM-based speech synthesis. In *Proceedings of the 8th ISCA Workshop on Speech Synthesis (SSW)*, pages 83–87, Barcelona, Spain.

Villavicencio, F., Bonada, J., Yamagishi, J., and Pucher, M. (2015). Efficient pitch estimation on natural opera-singing by a spectral correlation based strategy. *IEICE Technical report*, 115(146).

Bibliography

Akira, H., Haider, F., Cerrato, L., Campbell, N., and Luz, S. (2015). Detection of cognitive states and their correlation to speech recognition performance in speech-to-speech machine translation systems. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 2539–2543.

Allen, J. (1995). *Natural language understanding*. Benjamin/Cummings.

Amunts, K., Schleicher, A., Bürgel, U., Mohlberg, H., Uylings, H., and Zilles, K. (1999). Broca's region revisited: cytoarchitecture and intersubject variability. *Journal of Comparative Neurology*, 412(2):319–341.

Andrade-Miranda, G., Bernardoni, N. H., and Godino-Llorente, J. I. (2015). A new technique for assessing glottal dynamics in speech and singing by means of optical-flow computation. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 2182–2186.

Appel, R. and Beerends, J. G. (2002). On the quality of hearing one's own voice. *Journal of the Audio Engineering Society*, 50(4):237–248.

Arisoy, E., Sainath, T. N., Kingsbury, B., and Ramabhadran, B. (2012). Deep neural network language models. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 20–28. Association for Computational Linguistics.

Asaei, A., Cernak, M., and Bourlard, H. (2015). On compressibility of neural network phonological features for low bit rate speech coding. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 418–422.

Ashby, S., Barbosa, S., and Ferreira, J. P. (2010). Introducing non-standard Luso-African varieties into the digital domain. In *Proceedings of FALA 2010 - VI Jornadas en Tecnologa del Habla and II Iberian SLTech Workshop*, pages 247–250.

Astrinaki, M., Yamagishi, J., King, S., D'Alessandro, N., and Dutoit, T. (2013). Reactive accent interpolation through an interactive map application. In Bimbot, F., Cerisara, C., Fougeron, C., Gravier, G., Lamel, L., Pellegrino, F., and Perrier, P., editors, *Proceedings of the 14th Conference of the International Speech Communication Association (Interspeech 2013)*, pages 1877–1878, Lyon, France. ISCA.

Atti, V., Sinder, D. J., Subasingha, S., Rajendran, V., Dewasurendra, D., Chebiyyam, V., Varga, I., Krishnan, V., Schubert, B., Lecomte, J., et al. (2015). Improved error resilience for VOLTE and VOIP with 3GPP EVS channel aware coding. In *Acoustics,*

Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pages 5713–5717. IEEE.

Bahl, L. R., Jelinek, F., and Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 5(2):179–190.

Bailly, G., Bérar, M., Elisei, F., and Odisio, M. (2003). Audiovisual speech synthesis. *International Journal of Speech Technology*, 6:331–346.

Bailly, G., Govokhina, O., Elisei, F., and Breton, G. (2009). Lip-synching using speaker-specific articulation, shape and appearance models. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009(769494):1–11.

Bailly, G., Perrier, P., and Vatikiotis-Bateson, E. (2012). *Audiovisual speech processing*. Cambridge University Press.

Baker, J. K. (1975). The DRAGON system - An overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):24–29.

Barouti, M., Papadopoulos, K., and Kouroupetroglou, G. (2013). Synthetic and natural speech intelligibility in individuals with visual impairments: Effects of experience and presentation rate. In *European AAATE Conference*, pages 695–699, Portugal.

Beigi, H. (2011). *Fundamentals of speaker recognition*. Springer Science & Business Media.

Bellegarda, J. (2000). Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE*, 88(8):1279–1296.

Bellegarda, J. R. (2014). Spoken language understanding for natural interaction: The Siri experience. In *Natural Interaction with Robots, Knowbots and Smartphones*, pages 3–14. Springer.

Benesty, J., Sondhi, M. M., and Huang, Y. (2007). *Springer handbook of speech processing*. Springer Science & Business Media.

Berger, M., Hofer, G., and Shimodaira, H. (2011). Carnival - Combining speech technology and computer animation. *Computer Graphics and Applications, IEEE*, 31(5):80–89.

Bimbot, F., Bonastre, J. F., Fredouille, C., Gravier, G., Magrin-Chagnolleaua, I., Meignier, S., Merlin, T., Ortega-Garcia, J., and Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. *EURASIP J. Applied Signal Process.*, 4:430–451.

Black, A., Zen, H., and Tokuda, K. (2007). Statistical parametric speech synthesis. In *Proc. ICASSP 2007*, pages 1229–1232.

Blackburn, P. and Bos, J. (2005). Representation and inference for natural language. *A first course in computational semantics. CSLI.*

Bo, L.-J. (2000). Forensic voice identification in France. *Speech Commun.*, 31(2-3):205 – 224.

Bőhm, T. and Shattuck-Hufnagel, S. (2007). Utterance-final glottalization as a cue for familiar speaker recognition. In *Proc. Interspeech, Antwerp*, pages 2657–2660.

Bonastre, J.-F., Matrouf, D., and Fredouille, C. (2006). Transfer function-based voice transformation for speaker recognition. In *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey)*, pages 1–6.

Bonastre, J.-F., Matrouf, D., and Fredouille, C. (2007). Artificial impostor voice transformation effects on false acceptance rates. In *Proc. Int. Speech Commun. Association (Interspeech)*, pages 2053–2056.

Bregler, C., Covell, M., and Slaney, M. (1997). Video rewrite: driving visual speech with audio. In *Proc. SIGGRAPH*, pages 353–360, Los Angeles, CA, USA.

Budanitsky, A. and Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, 2nd Meeting of the North American Chapter of the ACL*, pages 29–34, Pittsburgh, USA.

Cakmak, H., Urbain, J., and Dutoit, T. (2015). Synchronization rules for HMM-based audio-visual laughter synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 2304–2308. IEEE.

Cakmak, H., Urbain, J., Dutoit, T., and Tilmanne, J. (2014a). The AV-LASYN database: A synchronous corpus of audio and 3D facial marker data for audio-visual laughter synthesis. In *LREC*, pages 3398–3403.

Cakmak, H., Urbain, J., Tilmanne, J., and Dutoit, T. (2014b). Evaluation of HMM-based visual laughter synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4578–4582. IEEE.

Campbell, N. (2006). Conversational speech synthesis and the need for some laughter. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1171–1178.

Campbell, W. M., Sturim, D. E., and Reynolds, D. A. (2006). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Process. Lett.*, 13(5):308–311.

Cernak, M., Potard, B., and Garner, P. N. (2015). Phonological vocoding using artificial neural networks. In *IEEE 40th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Chang, M.-W., Ratinov, L.-A., Rizzolo, N., and Roth, D. (2008). Learning and inference with constraints. In *AAAI*, pages 1513–1518.

Chelba, C., Engle, D., Jelinek, F., Jimenez, V. M., Khudanpur, S., Mangu, L., Printz, H., Ristad, E., Rosenfeld, R., Stolcke, A., and Wu, D. (1997). Structure and performance of a dependency language model. In *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97)*, volume 5, pages 2775–2778, Rhodes, Greece.

Chelba, C. and Jelinek, F. (1998). Exploiting syntactic structure for language modeling. In *Proceedings of the Conference of the Association for Computational Linguistics (COLING-ACL'98)*, pages 225–231, Montreal, Canada.

Chen, N., Qian, Y., Dinkel, H., Chen, B., and Yu, K. (2015). Robust deep feature for spoofing detection - the SJTU system for ASVspoof 2015 challenge. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Chen, S. H., Hsieh, C. H., Chiang, C. Y., Hsiao, H. C., Wang, Y. R., Liao, Y. F., and Yu, H. M. (2014). Modeling of speaking rate influences on Mandarin speech prosody and its application to speaking rate-controlled TTS. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(7):1158–1171.

Choi, J., Kim, J., Kang, S. J., and Kim, N. S. (2015). Reverberation-robust acoustic indoor localization. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 3422–3425.

Chu, W. C. (2004). *Speech coding algorithms: foundation and evolution of standardized coders*. John Wiley & Sons.

Clark, J. E. and Yallop, C. (2000). *An introduction to phonetics and phonology*. Foreign Language Teaching and Research Press.

Clyne, M. (1991). *Pluricentric languages: differing norms in different nations*, volume 62. Walter de Gruyter.

Cohen, M. and Massaro, D. (1993). Modeling coarticulation in synthetic visual speech. In Magnenat-Thalmann, N. and Thalmann, D., editors, *Models and Techniques in Computer Animation*, pages 139–156. Springer-Verlag.

Cohen, M. H., Giangola, J. P., and Balogh, J., editors (2004). *Voice User Interface Design*. Addison-Wesley.

Cook, S. A. (1971). The complexity of theorem-proving procedures. In *Proceedings of the 3rd annual ACM symposium on Theory of computing*, pages 151–158, Ohio, USA.

Covell, M., Withgott, M., and Slaney, M. (1998). Mach1: Nonuniform time-scale modification of speech. In *Proc. ICASSP*, volume 1, pages 349–352, Seattle, USA. IEEE.

Creer, S., Green, P., and Cunningham, S. (2009). Voice banking. *Advances in clinical neuroscience & rehabilitation*, 9(2):16 – 18.

Crevier, D. and Lepage, R. (1997). Knowledge-based image understanding systems: A survey. *Computer Vision and Image Understanding*, 67(2):161 – 185.

Csapó, T. G. and Lulich, S. M. (2015). Error analysis of extracted tongue contours from 2D ultrasound images. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 2157–2161.

De Leon, P. L. and Stewart, B. (2013). Synthetic speech detection based on selectedword discriminators. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3004–3008. IEEE.

De Leon, P. L., Stewart, B., and Yamagishi, J. (2012). Synthetic speech discrimination using pitch pattern statistics derived from image analysis. In *INTERSPEECH*, pages 370–373.

Dehaene-Lambertz, G., Dehaene, S., and Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science*, 298(5600):2013–2015.

Demetriou, G., Atwell, E., and Souter, C. (1997). Large-scale lexical semantics for speech recognition support. In *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97)*, pages 2755–2758, Rhodes, Greece.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., and Brumberg, J. S. (2010). Silent speech interfaces. *Speech Communication*, 52(4):270–287.

Deng, Z. and Neumann, U. (2006). eFASE: expressive facial animation synthesis and editing with phoneme-isomap controls. In *Proc. Eurographics SCA*, pages 251–260, Aire-la-Ville, Switzerland.

Deng, Z. and Neumann, U. (2008). *Data-driven 3D facial animation*. Springer.

Dietz, M., Multrus, M., Eksler, V., Malenovsky, V., Norvell, E., Pobloth, H., Miao, L., Wang, Z., Laaksonen, L., Vasilache, A., Kamamoto, Y., Kikuiri, K., Ragot, S., Faure, J., Ehara, H., Rajendran, V., Atti, V., Sung, H., Oh, E., Yuan, H., and Zhu, C. (2015). Overview of the EVS codec architecture. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5698–5702.

Disch, S., Neukam, C., and Schmidt, K. (2015). Temporal tile shaping for spectral gap filling in audio transform coding in EVS. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5873–5877.

Dressler, W. U. and Wodak, R. (1982). Sociophonological methods in the study of sociolinguistic variation in Viennese German. *Language in Society 11*, 11:339–370.

Dupont, S. and Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *Multimedia, IEEE Transactions on*, 2(3):141–151.

Eksler, V., Jelinek, M., and Salami, R. (2015). Efficient handling of mode switching and speech transitions in the EVS codec. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5918–5921.

Emami, A. (2015). Efficient machine translation decoding with slow language models. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 2376–2379.

Espi, M., Fujimoto, M., Kinoshita, K., and Nakatani, T. (2015). Feature extraction strategies in deep learning based acoustic event detection. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 2922–2926.

Evans, N., Yamagishi, J., and Kinnunen, T. (2013a). Spoofing and countermeasures for speaker verification: a need for standard corpora, protocols and metrics. `http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2013-05/spoofing/`.

Evans, N. W., Kinnunen, T., and Yamagishi, J. (2013b). Spoofing and countermeasures for automatic speaker verification. In *INTERSPEECH*, pages 925–929.

Ezzat, T., Geiger, G., and Poggio, T. (2002). Trainable videorealistic speech animation. In *Proc. SIGGRAPH*, pages 388–398, New York, NY, USA.

Facewaretech (2015). Facewaretech. `http://facewaretech.com/`.

Farrus, M., Erro, D., and Hern, J. (2008). Speaker recognition robustness to voice conversion. *IV Jornadas de Reconocimiento Biometrico de Personas*, pages 73–82.

Fernyhough, C. and Russell, J. (1997). Distinguishing one's own voice from those of others: A function for private speech? *International Journal of Behavioral Development*, 20(4):651–665.

Festival (2015). The Festival speech synthesis system. `http://www.cstr.ed.ac.uk/projects/festival/`.

Flanagan, J. L., Coker, C. H., Rabiner, L. R., Schafer, R. W., and Umeda, N. (1970). Synthetic voices for computers. *IEEE Spectrum*, 7(10):22–45.

Fraser, M. and King, S. (2007). The Blizzard Challenge 2007. In *Proc. Blizzard 2007 (in Proc. Sixth ISCA Workshop on Speech Synthesis)*, Bonn, Germany.

Furui, S. (2000). Digital speech processing, synthesis, and recognition (revised and expanded). *Digital Speech Processing, Synthesis, and Recognition (Second Edition, Revised and Expanded)*.

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *J. Acoust. Soc. Am.*, 63(1):223–230.

Ge, S. S., Khatib, O., Cabibihan, J.-J., Simmons, R., and Williams, M. A. (2012). *Social Robotics: 4th International Conference, ICSR 2012, Chengdu, China, October 29-31, 2012, Proceedings*, volume 7621. Springer.

Genoud, D. and Chollet, G. (1998). Speech pre-processing against intentional imposture in speaker recognition. In *Proc. Int. Conf. Spoken Language Process. (ICSLP)*, volume 2, pages 105–108.

Gick, B., Wilson, I., and Derrick, D. (2012). *Articulatory phonetics*. John Wiley & Sons.

Gildea, D. and Hofmann, T. (1999). Topic-based language models using EM. In *Proceedings of 6th European Conference On Speech Communication and Technology (Eurospeech'99)*, pages 2167–2170, Budapest, Hungary.

Goel, N., Thomas, S., Agarwal, M., Akyazi, P., Burget, L., Feng, K., Ghoshal, A., Glembek, O., Karafiat, M., Povey, D., Rastrow, A., Rose, R. C., and Schwarz, P. (2010). Approaches to automatic lexicon learning with limited training examples. In *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5094–5097, Dallas, Texas, USA.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. Academic Press, London.

Goodrich, M. A. and Schultz, A. C. (2007). Human-robot interaction: a survey. *Foundations and trends in human-computer interaction*, 1(3):203–275.

Google (2015). Google Glass. `http://www.google.com/glass/start/`.

Govokhina, O., Bailly, G., and Breton, G. (2007). Learning optimal audiovisual phasing for a HMM-based control model for facial animation. In *Proc. SSW6*, pages 1–4, Bonn, Germany.

Greisbach, R. (1992). Reading aloud at maximal speed. *Speech Comm.*, 11(4-5):469 – 473.

Grohe, A.-K., Poarch, G. J., Hanulíková, A., and Weber, A. (2015). Production inconsistencies delay adaptation to foreign accents. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 3115–3119.

Hall, T. A. (2011). *Phonologie: Eine Einführung*. Walter de Gruyter.

He, L. and Gupta, A. (2001). Exploring benefits of non-linear time compression. In *Proc. ACM Int. Conf. on Multimedia*, pages 382–391, Ottawa, Canada. ACM.

Heeman, P. (1998). POS tagging versus classes in language modeling. In *Proceedings of the 6th Workshop on Very Large Corpora*, pages 179–187, Montreal, Canada.

Hempel, T. (2008). *Usability of Speech Dialog Systems: Listening to the Target Audience*. Springer Science & Business Media.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97.

Hofer, G. and Richmond, K. (2010). Comparison of HMM and TMDN methods for lip synchronisation. In *Proc. INTERSPEECH*, pages 454–457, Makuhari, Japan.

Hofer, G., Yamagishi, J., and Shimodaira, H. (2008). Speech-driven lip motion generation with a trajectory HMM. In *Proc. INTERSPEECH*, pages 2314–2317, Brisbane, Australia.

Hsieh, C.-H., Chiang, C.-Y., Wang, Y.-R., Yu, H.-M., Chen, S.-H., et al. (2012). A new approach of speaking rate modeling for Mandarin speech prosody. In *13th Annual Conference of the International Speech Communication Association 2012 (INTERSPEECH 2012)*, pages 654–657.

HTS (2015). Hmm-based speech synthesis system (hts) - extensions. `http://hts.sp.nitech.ac.jp/Extensions`.

Huang, C., Chang, E., Zhou, J., and Lee, K.-F. (2000). Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition. In *INTERSPEECH*, pages 818–821. Citeseer.

Huang, C., Chen, T., and Chang, E. (2004). Accent issues in large vocabulary continuous speech recognition. *International Journal of Speech Technology*, 7(2-3):141–153.

Huang, X., Acero, A., Hon, H.-W., and Foreword By-Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice Hall PTR.

Hugdahl, K., Ek, M., Takio, F., Rintee, T., Tuomainen, J., Haarala, C., and Hmlinen, H. (2004). Blind individuals show enhanced perceptual and attentional sensitivity for identification of speech sounds. *Cognitive Brain Research*, 19(1):28 – 32.

Humphries, J. J., Woodland, P. C., and Pearce, D. (1996). Using accent-specific pronunciation modelling for robust speech recognition. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 4, pages 2324–2327. IEEE.

Hunt, A. J. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 373–376. IEEE.

ISCA (2015). Interspeech 2015. `http://interspeech2015.org/papers/scientific-areas/`.

Isogai, J., Yamagishi, J., and Kobayashi, T. (2005). Model adaptation and adaptive training using ESAT algorithm for HMM-based speech synthesis. In *Proc. EUROSPEECH 2005*, pages 2597–2600.

Iwano, K., Yamada, M., Togawa, T., , and Furui, S. (2002). Speech-rate-variable HMM-based Japanese TTS system. In *Proc. TTS2002*.

Janse, E. (2004). Word perception in fast speech: Artificially time-compressed vs. naturally produced fast speech. *Speech Comm.*, 42(2):155–173.

Janse, E., Nooteboom, S., and Quené, H. (2003). Word-level intelligibility of time-compressed speech: Prosodic and segmental factors. *Speech Comm.*, 41(2):287–301.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–557.

Jelinek, F. (1997). *Statistical methods for speech recognition*. MIT press.

Johnson, K. (2011). *Acoustic and auditory phonetics*. John Wiley & Sons.

Jokinen, E., Lecomte, J., Schinkel-Bielefeld, N., and Backstrom, T. (2015). Intelligibility evaluation of speech coding standards in severe background noise and packet loss conditions. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5152–5156.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing*. Prentice Hall.

Karaiskos, V., King, S., Clark, R. A. J., and Mayo, C. (2008). The Blizzard challenge 2008. In *Proc. Blizzard Challenge Workshop*, Brisbane, Australia.

Karhila, R. and Wester, M. (2011). Rapid adaptation of foreign-accented HMM-based speech synthesis. In *Proceedings of INTERSPEECH*, pages 2801–2804. ISCA.

Kat, L. and Fung, P. (1999). Fast accent identification and accented speech recognition. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 1, pages 221–224. IEEE.

Kim, T., Yue, Y., Taylor, S., and Matthews, I. (2015). A decision tree framework for spatiotemporal sequence prediction. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 577–586. ACM.

King, S., Tokuda, K., Zen, H., and Yamagishi, J. (2008). Unsupervised adaptation for HMM-based speech synthesis. In *Proceedings of INTERSPEECH*, pages 1869–1872, Brisbane, Australia.

Kinnunen, T., Karpov, E., and Franti, P. (2006). Real-time speaker identification and verification. *IEEE Trans. Audio, Speech, and Language Process.*, 14(1):277–288.

Kinnunen, T., Wu, Z., Evans, N., and Yamagishi, J. (2015). ASVspoof 2015: Automatic speaker verification spoofing and countermeasures challenge. `http://www.spoofingchallenge.org`.

Kinoshita, K., Delcroix, M., Ogawa, A., and Nakatani, T. (2015). Text-informed speech enhancement with deep neural networks. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 1760–1764.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Kolluru, B., Wan, V., Latorre, J., Yanagisawa, K., and Gales, M. J. (2014). Generating multiple-accent pronunciations for TTS using joint sequence model interpolation. In *INTERSPEECH*, pages 1273–1277.

Kondoz, A. M. (2005). *Digital speech: coding for low bit rate communication systems*. John Wiley & Sons.

Krenn, B., Endrass, B., Kistler, F., and André, E. (2014). Effects of language variety on personality perception in embodied conversational agents. In *Human-Computer Interaction. Advanced Interaction Modalities and Techniques*, pages 429–439. Springer.

Krenn, B., Schreitter, S., Neubarth, F., and Sieber, G. (2012). Social evaluation of artificial agents by language varieties. In *Intelligent Virtual Agents*, pages 377–389. Springer.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11):831–843.

Labov, W. (1972). *Sociolinguistic patterns*. Oxford, Blackwell.

Ladefoged, P. and Johnson, K. (2014). *A course in phonetics*. Cengage learning.

Lalanne, D., Nigay, L., Palanque, p., Robinson, P., Vanderdonckt, J., and Ladry, J.-F. (2009). Fusion engines for multimodal input: A survey. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI '09, pages 153–160, New York, NY, USA. ACM.

Lancker, D. V. and Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, 25(5):829–834.

Lau, Y. W., Wagner, M., and Tran, D. (2004). Vulnerability of speaker verification to voice mimicking. In *Proc. Int. Symp. Intelligent Multimedia, Video and Speech Process.*, pages 145 – 148.

Lebeter, J. and Saunders, S. (2010). The effects of time compression on the comprehension of natural and synthetic speech. *Working Papers of the Linguistics Circle*, 20(1):63–81.

Lecumberri, M. L. G., Barra-Chicote, R., Ramón, R. P., Yamagishi, J., and Cooke, M. (2014). Generating segmental foreign accent. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1302–1306.

Lee, C.-H., Soong, F. K., and Paliwal, K. (2012). *Automatic speech and speaker recognition: advanced topics*, volume 355. Springer Science & Business Media.

Lee, C. M., Shin, J. W., and Kim, N. S. (2015). DNN-based residual echo suppression. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 1730–1734.

Lei, M., Yamagishi, J., Richmond, K., Ling, Z.-H., King, S., and Dai, L.-R. (2011). Formant-controlled HMM-based speech synthesis. In *Proc. Interspeech*, pages 2777–2780, Florence, Italy.

Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Khudanpur, S., Schwartz, L., Thornton, W. N., Weese, J., and Zaidan, O. F. (2009). Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139. Association for Computational Linguistics.

Liang, H. and Dines, J. (2011). Phonological knowledge guided HMM state mapping for cross-lingual speaker adaptation. In *Proceedings of INTERSPEECH*, pages 1825–1828.

Liang, H., Dines, J., and Saheer, L. (2010). A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for HMM-based speech synthesis. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4598–4601. IEEE.

Liberman, A. M. (1970). Some characteristics of perception in the speech mode. *Perception and its disorders*, XLVIII(11).

Lindberg, J. and Blomberg, M. (1999). Vulnerability in speaker verification – a study of possible technical imposter techniques. In *Proc. European Conf. Speech Communication and Technology (Eurospeech)*, volume 3, pages 1211–1214.

Ling, Z.-H., Richmond, K., Yamagishi, J., and Wang, R.-H. (2008). Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge. In *Proc. Interspeech*, pages 573–576, Brisbane, Australia.

Ling, Z.-H., Richmond, K., Yamagishi, J., and Wang, R.-H. (2009). Integrating articulatory features into HMM-based parametric speech synthesis. *Trans. Audio, Speech, and Language Processing*, 17(6):1171–1185.

Ling, Z.-H. and Wang, R.-H. (2006). HMM-based unit selection using frame sized speech segments. In *INTERSPEECH*, pages 2034–2037.

Liu, C., Xu, P., and Sarikaya, R. (2015a). Deep contextual language understanding in spoken dialogue systems. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Liu, Y., Tian, Y., He, L., Liu, J., and Johnson, M. T. (2015b). Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Longworth, C. and Gales, M. (2009). Combining derivative and parametric kernels for speaker verification. *IEEE Trans. Audio, Speech, and Language Process.*, 17(4):748–757.

Loots, L. and Niesler, T. (2011). Automatic conversion between pronunciations of different English accents. *Speech Communication*, 53(1):75–84.

Loweimi, E., Barker, J., and Hain, T. (2015). Source-filter separation of speech signal in the phase domain. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 598–602.

Lu, H.-t., Liou, Y.-m., Lee, H.-y., and Lee, L.-s. (2015). Semantic retrieval of personal photos using a deep autoencoder fusing visual features with speech annotations represented as word/paragraph vectors. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 140–144.

Manning, C. D., Raghavan, P., Schütze, H., et al. (2008). *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*, volume 999. MIT Press.

Marcel, S. (2014). Trusted biometrics under spoofing attacks (TABULA RASA). `http://www.tabularasa-euproject.org`.

Massimino, P. (2005). From marked text to mixed speech and sound. In *ICAD 2005 workshop Combining Speech and Sound in the User Interface*.

Masuko, T., Hitotsumatsu, T., Tokuda, K., and Kobayashi, T. (1999). On the security of HMM-based speaker verification systems against imposture using synthetic speech. In *Proc. European Conf. Speech Communication and Technology (Eurospeech)*, pages 1223–1226.

Masuko, T., Kobayashi, T., Tamura, M., Masubuchi, J., and Tokuda, K. (1998). Text-to-visual speech synthesis based on parameter generation from HMM. In *Proc. ICASSP*, volume 6, pages 3745 –3748 vol.6.

Masuko, T., Tokuda, K., and Kobayashi, T. (2000). Imposture using synthetic speech against speaker verification based on spectrum and pitch. In *Proc. Int. Conf. Spoken Language Process. (ICSLP)*, volume 2, pages 302–305.

Masuko, T., Tokuda, K., Kobayashi, T., and Imai, S. (1996). Speech synthesis using HMMs with dynamic features. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 389–392.

Masuko, T., Tokuda, K., Kobayashi, T., and Imai, S. (1997). Voice characteristics conversion for HMM-based speech synthesis system. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 1611–1614.

Matrouf, D., Bonastre, J.-F., and Fredouille, C. (2006). Effect of speech transformation on impostor acceptance. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, volume 2, pages 933–936.

Matsui, T. and Furui, S. (1995). Likelihood normalization for speaker verification using a phoneme- and speaker-independent model. *Speech Commun.*, 17(1-2):109–116.

Mattheyses, W. and Verhelst, W. (2015). Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66:182–217.

McClanahan, R. D., Stewart, B., and De Leon, P. L. (2014). Performance of i-vector speaker verification and the detection of synthetic speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3779–3783. IEEE.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*.

Mendels, G., Cooper, E., Soto, V., Hirschberg, J., Gales, M., Knill, K., Ragni, A., and Wang, H. (2015). Improving speech recognition and keyword search for low resource languages using web data. *Proc. Interspeech, Dresden, Germany*, pages 829–833.

Montague, R. (1973). *The proper treatment of quantification in ordinary English*. Springer.

Moos, A. and Trouvain, J. (2007). Comprehension of ultra-fast speech – blind vs. 'normally hearing' persons. In *Proc. Int. Congress of Phonetic Sciences*, volume 1, pages 677–680.

Moosmüller, S. (1987). *Soziophonologische Variation im gegenwärtigen Wiener Deutsch*. Franz Steiner Verlag, Stuttgart.

Moosmüller, S. (1991). *Hochsprache und Dialekt in Österreich. Soziophonologische Untersuchungen zu ihrer Abgrenzung in Wien, Graz, Salzburg und Innsbruck*. Wien: Böhlau.

Moosmüller, S. (1997). Evaluation of language use in public discourse: language attitudes in austria. *Stevenson (ed.)*, pages 259–80.

Moosmüller, S. (2012). The roles of stereotypes, phonetic knowledge, and phonological knowledge in the evaluation of dialect authenticity. *Proceedings ofSociophonetics, at the Crossroads of Speech Variation, Processing and Communication. Edizioni della Normale, Pisa*, pages 49–52.

Moosmüller, S., Schmid, C., and Brandstätter, J. (in press). Standard Austrian German. *Journal of the International Phonetic Association*.

Moosmüller, S. and Vollmann, R. (2001). Natürliches Driften im Lautwandel: Die Monophthongierung im österreichischen Deutsch. *Zeitschrift für Sprachwissenschaft*, 20/I:42–65.

Morgan, D. P. and Scofield, C. L. (1991). *Neural networks and speech processing*. Springer.

Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [from the field]. *Robotics & Automation Magazine, IEEE*, 19(2):98–100.

Mowlaee, P., Saeidi, R., and Stylanou, Y. (2014). Interspeech 2014 special session: Phase importance in speech processing applications. In *Proceedings of the 15th International Conference on Spoken Language Processing*.

Nagisetty, S., Liu, Z., Kawashima, T., Ehara, H., Zhou, X., Wang, B., Liu, Z., Miao, L., Gibbs, J., Laaksonen, L., Atti, V., Rajendran, V., Krishnan, V., Sung, H., and Choo, K. (2015). Low bit rate high-quality MDCT audio coding of the 3GPP EVS standard. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5883–5887.

Nass, C. and Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3):171.

Navas, E., Hernaez, I., Erro, D., Salaberria, J., Oyharçabal, B., and Padilla, M. (2014). Developing a basque tts for the navarro-lapurdian dialect. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 11–20. Springer.

Newman, R. S. and Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics*, 35(1):85 – 103.

Nguyen, T. T. T., d'Alessandro, C., Rilliard, A., and Tran, D. D. (2013). HMM-based TTS for Hanoi Vietnamese: issues in design and evaluation. In *Proceedings of the 14th Conference of the International Speech Communication Association (Interspeech 2013)*, pages 2311–2315, Lyon, France.

Nilsson, N. J. (2010). *The quest for Artificial Intelligence*. Cambridge University Press.

Nygaard, L. C. and Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & psychophysics*, 60(3):355–376.

Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5(1):42–46.

Ooigawa, T. (2015). Perception of italian liquids by japanese listeners: Comparisons to spanish liquids. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 3135–3139.

Ordin, M. and Polyanskaya, L. (2015). Acquisition of english speech rhythm by monolingual children. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 3120–3124.

O'Reilly, C., Marples, N. M., Kelly, D. J., and Harte, N. (2015). Quantifying difference in vocalizations of bird populations. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 3417–3421.

ORF (2015). Lohner als ÖBB-Computerstimme. `http://wien.orf.at/news/stories/2745538/`.

O'Shaughnessy, D. (1987). *Speech communication: Human and machine*. Universities press.

Oura, K., Yamagishi, J., Wester, M., King, S., and Tokuda, K. (2012). Analysis of unsupervised cross-lingual speaker adaptation for hmm-based speech synthesis using kld-based transform mapping. *Speech Communication*, 54(6):703 – 714.

Oviatt, S. (2003). Multimodal interfaces. *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*, 14:286–304.

Oviatt, S. and Cohen, P. (2000). Perceptual user interfaces: Multimodal interfaces that process what comes naturally. *Commun. ACM*, 43(3):45–53.

Papadopoulos, K., Argyropoulos, V. S., and Kouroupetroglou, G. (2008). Discrimination and comprehension of synthetic speech by students with visual impairments: The case of similar acoustic patterns. *Journal of Visual Impairment & Blindness*, 102(7):420–429.

Parke, F. I. and Waters, K. (2008). *Computer facial animation*. CRC Press.

Pellom, B. L. and Hansen, J. H. L. (1999). An experimental study of speaker verification sensitivity to computer voice-altered imposters. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 837–840.

Pfister, B. and Kaufmann, T. (2008). *Sprachverarbeitung: grundlagen und methoden der sprachsynthese und spracherkennung*. Springer-Verlag.

Phan, H., Hertel, L., Maass, M., Mazur, R., and Mertins, A. (2015). Representing nonspeech audio signals through speech classification models. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 3441–3445.

Picart, B., Drugman, T., and Dutoit, T. (2011). Continuous control of the degree of articulation in HMM-based speech synthesis. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1797–1800, Florence, Italy.

Picart, B., Drugman, T., and Dutoit, T. (2014). Hmm-based speech synthesis with various degrees of articulation: A perceptual study. *Neurocomputing*, 132:142 – 147. Innovations in Nature Inspired Optimization and Learning MethodsMachines learning for Non-Linear ProcessingSelected papers from the Third World Congress on Nature and Biologically Inspired Computing (NaBIC2011)Selected papers from the 2011 International Conference on Non-Linear Speech Processing (NoLISP 2011).

Pisoni, D. and Remez, R. (2008). *The Handbook of Speech Perception*. John Wiley & Sons.

Pisoni, D. B. (1997). Perception of synthetic speech. In *Progress in speech synthesis*, pages 541–560. Springer.

Port, R. F. (1981). Linguistic timing factors in combination. *J. Acoust. Soc. Am.*, 69(1):262–274.

Qian, Y., Xu, J., and Soong, F. K. (2011). A frame mapping based hmm approach to cross-lingual voice transformation. In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5120–5123, Prague, Czech Republic.

Qin, L., Ling, Z., Wu, Y., Zhang, B., and Wang, R. (2006). HMM-based emotional speech synthesis using average emotion model. In *Proc. ISCSLP-2006 (Springer LNAI Book)*, pages 233–240.

Quatieri, T. F. (2002). *Speech signal processing*. Prentice Hall Signal Processing Series.

Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of speech recognition*. Prentice hall.

Rabiner, L., Rosenberg, A. E., and Levinson, S. E. (1978). Considerations in dynamic time warping algorithms for discrete word recognition. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 26:575–582.

Rabiner, L. R. and Schafer, R. W. (1978). *Digital processing of speech signals*. Prentice Hall.

Racca, D. N. and Jones, G. J. (2015). Incorporating prosodic prominence evidence into term weights for spoken content retrieval. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 1378–1382.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000a). Speaker verification using adapted Gaussian mixture models. *Dig. Sig. Process.*, 10:19–41.

Reynolds, M., Isaacs-Duvall, C., Sheward, B., and Rotter, M. (2000b). Examination of the effects of listening practice on synthesized speech comprehension. *Augmentative and Alternative Communication*, 16(4):250–259.

Richmond, K., Clark, R. A. J., and Fitt, S. (2010). On generating Combilex pronunciations via morphological analysis. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1974–1977, Makuhari, Japan.

Riedi, M. (1995). A neural-network-based model of segmental duration for speech synthesis. In *EUROSPEECH*.

Rosa, C., Lassonde, M., Pinard, C., Keenan, J. P., and Belin, P. (2008). Investigations of hemispheric specialization of self-voice recognition. *Brain and cognition*, 68(2):204–214.

Roth, D. and Yih, W.-t. (2005). Integer linear programming inference for conditional random fields. In *Proceedings of the 22nd international conference on Machine learning*, pages 736–743. ACM.

Ruiz, N., Gao, Q., Lewis, W., and Federico, M. (2015). Adapting machine translation models toward misrecognized speech with text-to-speech pronunciation rules and acoustic confusability. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 2247–2251.

Russell, M., DeMarco, A., Veaux, C., and Najafian, M. (2013). What's happening in accents & dialects? A review of the state of the art. In *UKSpeech Conference, Cambridge, 17/18 September 2013*, Cambridge, UK.

Russell, S. J. and Norvig, P. (1995). *Artificial Intelligence - A modern approach*. Prentice-Hall.

Saino, K., Zen, H., Nankaku, Y., Lee, A., and Tokuda, K. (2006). An HMM-based singing voice synthesis system. In *Proc. Interspeech*, pages 2274–2277, Pittsburgh, PA, USA.

Sako, S., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). HMM-based text-to-audio-visual speech synthesis. In *Proc. ICSLP*, pages 25–28, Beijing, China.

Sanchez, J., Saratxaga, I., Hernaez, I., Navas, E., and Erro, D. (2014). A cross-vocoder study of speaker independent synthetic speech detection using phase information. In *INTERSPEECH*, pages 1663–1667.

Sanchez, J., Saratxaga, I., Hernaez, I., Navas, E., and Erro, D. (2015). The aholab rps ssd spoofing challenge 2015 submission. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Sanderman, A. A. and Collier, R. (1997). Prosodic phrasing and comprehension. *Language and Speech*, 40(4):391–409.

Satoh, T., Masuko, T., Kobayashi, T., and Tokuda, K. (2001). A robust speaker verification system against imposture using an HMM-based speech synthesis system. In *Proc. European Conf. Speech Communication and Technology (Eurospeech)*, pages 759–762.

Saussure, F. D. (1983). *Course in General Linguistics*. Duckworth, London. (Original work published 1916).

Schönle, P. W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., and Conrad, B. (1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain Lang.*, 31:26–35.

Schuller, B. and Batliner, A. (2013). *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MüLler, C., and Narayanan, S. (2013). Paralinguistics in speech and languagestate-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39.

Schuller, B., Steidl, S., Batliner, A., Hantke, S., Hönig, F., Orozco-Arroyave, J. R., Nöth, E., Zhang, Y., and Weninger, F. (2015). The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, parkinsons & eating condition. In *Interspeech 2015*.

Serafini, L. and Bouquet, P. (2004). Comparing formal theories of context in ai. *Artificial intelligence*, 155(1):41–67.

Sizov, A., Khoury, E., Kinnunen, T., Wu, Z., and Marcel, S. (2015). Joint speaker verification and antispoofing in the-vector space. *Information Forensics and Security, IEEE Transactions on*, 10(4):821–832.

Souza, P., Gehani, N., Wright, R., and McCloy, D. (2013). The advantage of knowing the talker. *Journal of the American Academy of Audiology*, 24(8):689.

Spencer, A. (1995). *Phonology: theory and description*, volume 9. Wiley-Blackwell.

Sriskandaraja, K., Sethu, V., Le, P. N., and Ambikairajah, E. (2015). A model based voice activity detector for noisy environments. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 2297–2301.

Stan, A., Watts, O., Mamiya, Y., Giurgiu, M., Clark, R., Yamagishi, J., and King, S. (2013). TUNDRA: a multilingual corpus of found data for TTS research created with light supervision. In *Proceedings of the 14th Conference of the International Speech Communication Association (Interspeech 2013)*, pages 2331–2335, Lyon, France.

Stevens, K. (1997). Articulatory-acoustic-auditory relationships. In Hardcastle, W. J. and Laver, J., editors, *The handbook of phonetic sciences*, pages 462–506. Blackwell, Cambridge.

Stevens, K. N. (1998). Acoustic phonetics (vol. 30).

Sullivan, K. and Pelecanos, J. (2001). Revisiting carl bildt's impostor: Would a speaker verification system foil him? In Bigun, J. and Smeraldi, F., editors, *Audio- and Video-Based Biometric Person Authentication*, volume 2091 of *Lecture Notes in Computer Science*, pages 144–149. Springer Berlin / Heidelberg.

Syrdal, A. K., Bunnell, H. T., Hertz, S. R., Mishra, T., Spiegel, M. F., Bickley, C., Rekart, D., and Makashay, M. J. (2012). Text-to-speech intelligibility across speech rates. In *Proc. Interspeech*, Portland, USA.

Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T. (2005a). Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE Trans. Inf. Syst.*, E88-D(11):2484–2491.

Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T. (2005b). Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing,. *IEICE Trans. Inf. & Syst.*, E88-D(11):2484–2491.

Tamagawa, R., Watson, C. I., Kuo, I. H., MacDonald, B. A., and Broadbent, E. (2011). The effects of synthesized voice accents on user perceptions of robots. *International Journal of Social Robotics*, 3(3):253–262.

Tamura, M., Kondo, S., Masuko, T., and Kobayashi, T. (1999). Text-to-audio-visual speech synthesis based on parameter generation from HMM. In *Proc. EUROSPEECH*, pages 959–962, Budapest, Hungary.

Tamura, M., Masuko, T., Kobayashi, T., and Tokuda, K. (1998a). Visual speech synthesis based on parameter generation from HMM: Speech-driven and text-and-speech-driven approaches. In *Proc. AVSP*, pages 221–226.

Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T. (1998b). Speaker adaptation for HMM-based speech synthesis system using MLLR. In *The Third ESCA/COCOSDA Workshop on Speech Synthesis*, pages 273–276.

Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T. (2001). Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. In *Proc. ICASSP 2001*, pages 805–808.

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge university press.

Terry, L. (2011). *Audio-Visual Asynchrony Modeling and Analysis for Speech Alignment and Recognition*. Ph.d., Northwestern University, Chicago, IL.

Theobald, B., Bangham, J., Matthews, I., and Cawley, G. (2004). Near-videorealistic synthetic talking faces: implementation and evaluation. *Speech Communication*, 44(1-4):127–140.

Tokuda, K., Kobayashi, T., and Imai, S. (1995). Speech parameter generation from HMM using dynamic features. In *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 660–663, Detroit, MI, USA.

Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (1999). Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 229–232, Phoenix, AZ, USA.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000a). Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*, pages 1315–1318, Istanbul, Turkey.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000b). Speech parameter generation algorithms for hmm-based speech synthesis. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1315–1318. IEEE.

Traber, C. (1991). F0 generation with a data base of natural f0 patterns and with a neural network. In *The ESCA Workshop on Speech Synthesis*.

Trager, G. L. (1958). *Paralanguage: A first approximation*. University of Buffalo, Department of Anthropology and Linguistics.

Turing, A. M. (1950). *Computing machinery and intelligence*. Mind.

Turkle, S. (2007). Authenticity in the age of digital companions. *Interaction studies*, 8(3):501–517.

Urbain, J., Cakmak, H., and Dutoit, T. (2013). Evaluation of HMM-based laughter synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7835–7839. IEEE.

Van Eijck, J. and Unger, C. (2010). *Computational semantics with functional programming*. Cambridge University Press.

Van Lancker, D., Kreiman, J., and Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. part i: Recognition of backward voices. *Journal of phonetics*, 13:19–38.

Veaux, C., Yamagishi, J., and King, S. (2012). Using hmm-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders. In *INTERSPEECH*, pages 967–970.

Vilimek, R. and Hempel, T. (2005). Effects of speech and non-speech sounds on short-term memory and possible implications for in-vehicle use. In *ICAD 2005 workshop Combining Speech and Sound in the User Interface*.

Villalba, J., Miguel, A., Ortega, A., and Lleida, E. (2015). Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Wahlster, W. (2013). *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media.

Waibel, A., Jain, A., McNair, A., Saito, H., Hauptmann, A., and Tebelskis, J. (1991). JANUS: a speech-to-speech translation system using connectionist and symbolic processing strategies. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 793–796 vol.2.

Wang, L., Wu, Y.-J., Zhuang, X., and Soong, F. K. (2011). Synthesizing visual speech trajectory with minimum generation error. In *Proc. ICASSP*, pages 4580–4583.

Wang, L., Yoshida, Y., Kawakami, Y., and Nakagawa, S. (2015). Relative phase information for detecting human speech and spoofed speech. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Wang, Y.-Y., Deng, L., and Acero, A. (2005). Spoken language understandng: An introduction to the statistical framework. *IEEE Signal Processing Magazine*, 22(5):16–32.

Watson, C. I. and Marchi, A. (2014). Resources created for building new zealand english voices. In *Proceedings of the 15th Australasian International Conference on Speech Science and Technology*.

Watts, O., Stan, A., Clark, R., Mamiya, Y., Giurgiu, M., Yamagishi, J., and King, S. (2013). Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis. In *Proceedings of the 8th ISCA Workshop on Speech Synthesis (SSW)*, pages 101–106. ISCA.

Wester, M. and Karhila, R. (2011). Speaker similarity evaluation of foreign-accented speech synthesis using HMM-based speaker adaptation. In *Proc. ICASSP*, pages 5372–5375, Prague, Czech Republic.

Wong, C. H., Lee, T., Yeung, Y. T., and Ching, P. C. (2015). Modeling temporal dependency for robust estimation of LP model parameters in speech enhancement. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 1730–1734.

Wu, Y.-J., Nankaku, Y., and Tokuda, K. (2009a). State mapping based method for cross-lingual speaker adaptation in hmm-based speech synthesis. In *Proceedings of INTERSPEECH*, pages 528–531.

Wu, Y.-J., Nankaku, Y., and Tokuda, K. (2009b). State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 528–531, Brighton, United Kingdom.

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., and Li, H. (2015a). Spoofing and countermeasures for speaker verification: a survey. *Speech Communication*, 66:130–153.

Wu, Z., Khodabakhsh, A., Demiroglu, C., Yamagishi, J., Saito, D., Toda, T., and King, S. (2015b). Sas: A speaker verification spoofing database containing diverse attacks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4440–4444. IEEE.

Wu, Z., Larcher, A., Lee, K.-A., Chng, E., Kinnunen, T., and Li, H. (2013a). Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints. In *INTERSPEECH*, pages 950–954.

Wu, Z., Xiao, X., Chng, E. S., and Li, H. (2013b). Synthetic speech detection using temporal modulation feature. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7234–7238. IEEE.

Wutiwiwatchai, C., Thangthai, A., Chotimongkol, A., Hansakunbuntheung, C., and Thatphithakkul, N. (2011). Accent level adjustment in bilingual thai-english text-to-speech synthesis. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 295–299.

Yamagishi, J. and Kobayashi, T. (2007a). Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Transactions on Information and Systems*, E90-D(2):533–543.

Yamagishi, J. and Kobayashi, T. (2007b). Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Trans. Inf. & Syst.*, E90-D(2):533–543.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009a). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Trans. on Audio, Speech and Language Processing*, 17(1):66 –83.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009b). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Trans. Speech, Audio & Language Process.*, 17(1):66–83.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009c). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Trans. Audio, Speech, and Language Processing*, 17(1):66–83.

Yamagishi, J., Masuko, T., and Kobayashi, T. (2004). MLLR adaptation for hidden semi-Markov model based speech synthesis. In *Proc. ICSLP 2004*, pages 1213–1216.

Yamagishi, J., Nose, T., Zen, H., Ling, Z., Toda, T., Tokuda, K., King, S., and Renals, S. (2009d). Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 17(6):1208–1230.

Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S., and Renals, S. (2009e). A robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Trans. Speech, Audio & Language Process.*, 17(6):1208–1230.

Yamagishi, J., Ogata, K., Nakano, Y., Isogai, J., and Kobayashi, T. (2006). HSMM-based model adaptation algorithms for average-voice-based speech synthesis. In *Proc. ICASSP 2006*, pages 77–80.

Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., Hu, R., Oura, K., Tokuda, K., Karhila, R., and Kurimo, M. (2010). Thousands of voices for HMM-based speech synthesis – Analysis and application of TTS systems built on various ASR corpora. *IEEE Trans. Speech, Audio & Language Process.*, 18:984–1004.

Yamagishi, J., Veaux, C., King, S., and Renals, S. (2012). Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology*, 33(1):1–5.

Yamagishi, J. and Watts, O. (2010). The CSTR/EMIME HTS system for Blizzard challenge 2010. In *Proceedings of the Blizzard Challenge Workshop*, pages 1–6, Kansai Science City, Japan.

Yonan, C. A. and Sommers, M. S. (2000). The effects of talker familiarity on spoken word identification in younger and older listeners. *Psychology and aging*, 15(1):88.

Yoshimura, T., Masuko, T., Tokuda, K., Kobayashi, T., and Kitamura, T. (1997). Speaker interpolation in HMM-based speech synthesis system. In *Proc. Eurospeech*.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1999a). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. EUROSPEECH-99*, pages 2374–2350.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1999b). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 2374–2350, Budapest, Hungary.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2000a). Speaker interpolation for HMM-based speech synthesis system. *Acoustical Science and Technology*, 21(4):199–206.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2000b). Speaker interpolation for HMM-based speech synthesis system. *Journal of the Acoustical Science of Japan (E)*, 21(4):199–206.

Young, S. (2010). IEEE Signal Processing Magazine. *Cognitive User Interfaces*, 27:128–140.

Yu, D. and Deng, L. (2014). *Automatic Speech Recognition: A Deep Learning Approach*. Springer.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., and Tokuda, K. (2007a). The HMM-based speech synthesis system (HTS) version 2.0. In *Proc. SSW*, pages 294–299.

Zen, H., Oura, K., Nose, T., Yamagishi, J., Sako, S., Toda, T., Masuko, T., Black, A., and Tokuda, K. (2009a). Recent development of the HMM-based speech synthesis system (HTS). In *Proc. 2009 Asia-Pacific Signal and Information Processing Association (APSIPA)*, pages 121–130.

Zen, H., Senior, A., and Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7962–7966. IEEE.

Zen, H., Tokuda, K., and Black, A. W. (2009b). Statistical parametric speech synthesis. *Speech Commun.*, 51(11):1039–1064.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2004). Hidden semi-Markov model based speech synthesis. In *Proc. ICSLP 2004*, pages 1393–1396.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2007b). A hidden semi-markov model-based speech synthesis system. *IEICE Trans. Inf. Syst.*, E90-D(5):825–834.

Zhang, L., Karakos, D., Hartmann, W., Hsiao, R., Schwartz, R., and Tsakalidis, S. (2015). Enhancing low resource keyword spotting with automatically retrieved web documents. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 839–843.

Zhang, X.-L. and Wang, D. (2015). Multi-resolution stacking for speech separation based on boosted DNN. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 1745–1749.

Zhu, Z., Feng, X., and Yang, J. (2015). Lip movement and speech synchronization detection based on multimodal shift-invariant dictionary. In *2015 IEEE 16th International Conference on Communication Technology (ICCT)*, pages 768–772. IEEE.

Zissman, M. A. and Berkling, K. M. (2001). Automatic language identification. *Speech Communication*, 35(1):115–124.

A Sample webpages

Some links to sample pages from the principal references are not working anymore. Below you can find a list of all sample pages with updated links.

A.1 Text-to-Speech Synthesis

Principal ref.	[Toman et al., 2015]
Old URL	`http://userver.ftw.at/~mtoman/specom14/`
New URL	`http://mtoman.neuratec.com/thesis/interpolation/`
Principal ref.	[Valentini-Botinhao et al., 2015]
Old URL	`http://wiki.inf.ed.ac.uk/CSTR/SalbProject`
New URL	`http://wiki.inf.ed.ac.uk/CSTR/SalbProject`
Principal ref.	[Pucher et al., 2010b]
Old URL	`http://dialect-tts.ftw.at`
New URL	`http://web.archive.org/web/20150827100105/`
	`https://portal.ftw.at/projects/vsds`
Principal ref.	[Valentini-Botinhao et al., 2014]
Old URL	`http://wiki.inf.ed.ac.uk/CSTR/SalbProject`
New URL	`http://wiki.inf.ed.ac.uk/CSTR/SalbProject`

Table A.1: *New URLs for sample webpages.*

A.2 Audio-Visual Text-to-Speech Synthesis

Principal ref.	[Schabus et al., 2014]
Old URL	`http://userver.ftw.at/~schabus/jstsp2013`
New URL	`http://schabus.xyz/phd/audiovisual/`
Principal ref.	[Schabus et al., 2012]
Old URL	`http://userver.ftw.at/~schabus/interspeech2012/`
New URL	`http://schabus.xyz/phd/adaptation/`
Principal ref.	[Hollenstein et al., 2013]
Old URL	`http://userver.ftw.at/~schabus/avsp2013vc`
New URL	`http://schabus.xyz/phd/visualcontrol/plain.html`

B List of principal references

B.1 Text-to-Speech Synthesis

B.1.1 Journals

The 2010 paper on dialect interpolation in the *Speech Communication* journal has 34 citations according to Google Scholar, the very recent papers have clearly not yet been cited that often but will surely also be valuable for the research community. Because of this work I have been invited to give talks at a Dagstuhl workshop and at *National Institute of Informatics* (NII).

The ideas applied in Paper 1 were developed by me in the AMTV [Pucher, 2012a] project proposal. The paper was mainly written by Markus Toman, myself, and Sylvia Moosmüller. I was also co-supervising Markus Toman's PhD at FTW [Toman, 2016]. Basic ideas from Paper 2 and Paper 5 were conceived by me for the SALB [Pucher, 2013] project proposal. The ideas for Paper 3 were mine and it was written in large parts by me together with Dietmar Schabus. The whole voice training part was performed by myself. Paper 4 was also a result from my ideas from the SALB [Pucher, 2013] project, mainly written by myself. The experiments and paper writing for Paper 6 were mainly done by me.

1. Markus Toman, **Michael Pucher**, Sylvia Moosmüller, Dietmar Schabus, Unsupervised and phonologically controlled interpolation of Austrian German language varieties for speech synthesis. *Speech Communication*, Volume 72, 2015, pp. 176-193 [**Google scholar citations: 2, Journal impact factor: 1.256**].
2. Cassia Valentini-Botinhao, Markus Toman, **Michael Pucher**, Dietmar Schabus, Junichi Yamagishi, Intelligibility of time-compressed synthetic speech: compression method and speaking style. *Speech Communication*, Volume 74, pp. 52-64, November 2015 [**Google scholar citations: 0, Journal impact factor: 1.256**].
3. **Michael Pucher**, Dietmar Schabus, Junichi Yamagishi, Friedrich Neubarth, Volker Strom. Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis. *Speech Communication*, Volume 52, 2010, pp. 164-179 [**Google scholar citations: 34, Journal impact factor: 1.256**].

B.1.2 Conferences

4. **Michael Pucher**, Markus Toman, Dietmar Schabus, Cassia Valentini-Botinhao, Junichi Yamagishi, Bettina Zillinger, Erich Schmid. Influence of speaker familiarity on blind and visually impaired childrens perception of synthetic voices in audio games. In *Proceedings of INTERSPEECH 2015*, Dresden, Germany, pp. 1625-1629.
5. Cassia Valentini-Botinhao, Markus Toman, **Michael Pucher**, Dietmar Schabus, Junichi Yamagishi. Intelligibility analysis of fast synthesized speech. In *Proceedings of INTERSPEECH 2014*, Singapore, pp. 2922-2926.
6. **Michael Pucher**, Dietmar Schabus, Junichi Yamagishi. Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and sighted listeners. In *Proceedings of INTERSPEECH 2010*, Makuhari, Japan, pp. 2186-2189.

B.2 Audio-Visual Text-to-Speech Synthesis

The recent 2014 paper on joint audio-visual modeling in the *IEEE Journal of Selected Topics in Signal Processing* has 11 citations but will likely have an impact in the coming years.

Ideas from Paper 7-9 were developed by me and Gregor Hofer for the AVDS [Pucher, 2011] project proposal. I was also co-supervising Dietmar Schabus PhD at FTW [Schabus, 2014]. Experiments and algorithms for Paper 7 and 8 were conceived mainly by Dietmar Schabus and myself, and by me and Jakob Hollenstein for Paper 9.

B.2.1 Journals

7. Dietmar Schabus, **Michael Pucher**, Gregor Hofer. Joint Audiovisual Hidden Semi-Markov Model-based Speech Synthesis. *IEEE Journal of Selected Topics in Signal Processing*, Vol. 8, No. 2, pp. 336-347, April 2014 [**Google scholar citations: 11, Journal impact factor: 2.373**].

B.2.2 Conferences

8. Dietmar Schabus, **Michael Pucher**, Gregor Hofer. Speaker-adaptive visual speech synthesis in the HMM-framework. In *Proceedings of INTERSPEECH 2012*, Portland, USA, pp. 979-982.

9. Jakob Hollenstein, **Michael Pucher**, Dietmar Schabus. Visual Control of Hidden-Semi-Markov-Model based Acoustic Speech Synthesis. In *Proceedings of AVSP 2013*, Annency, France, pp. 31-35.

B.3 Speaker Verification Spoofing

B.3.1 Journals

The 2012 paper on speaker verification spoofing in the *IEEE Transactions on Audio, Speech, and Language Processing* has 57 citations. This journal paper and 3 conference papers received together 161 citations. These papers received much attention in the speaker verification community and led to a multi-disciplinary research topic of "voice anti-spoofing". Since the publication of our papers several initiatives have investigated the topic [Evans et al., 2013a; Marcel, 2014], with the latest initiative being this years spoofing challenge "ASVspoof 2015: Automatic Speaker Verification Spoofing and Countermeasures Challenge" at Interspeech 2015 [Kinnunen et al., 2015].

The original idea for the speaker verification spoofing topic was by me and Phillip L. De Leon, who was on a sabbatical at TU Vienna. The idea for phase-based detection was from Inma Hernaez, who did a sabbatical at FTW. For Paper 10 I wrote parts of the introduction, speech synthesis, and synthetic speech detection sections.

10. Phillip L. De Leon, **Michael Pucher**, Junichi Yamagishi, Inma Hernaez, Ibon Saratxaga. Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, Volume 20, Issue 8, October 2012, pp. 2280-2290 [**Google scholar citations: 57, Journal impact factor: 2.475**].

Mag. phil. Dipl.-Ing. Dr. techn. **Michael Pucher**

Senior Research Scientist
Acoustics Research Institute (ARI) of the Austrian
Academy of Sciences (ÖAW)
Wohllebengasse 12-14 / 1st Floor, Vienna A-1040, Austria
Tel.: +43/1/51581-2534 Mobile +43/664/8269861
Web: https://www.kfs.oeaw.ac.at
Email: michael.pucher@oeaw.ac.at

I obtained my doctoral degree (Dr.techn.) in Electrical and Information Engineering from Graz University of Technology in 2007. During the last years my work was focused on the improvement of state-of-the-art speech synthesis technologies for the synthesis of language varieties and audio-visual speech. I have also made significant contributions in the area of speaker verification spoofing, where we showed how adaptive synthesizers can spoof a speaker verification system. Currently I am working on conversion and modeling methods for accents of second language learners. This has many potential applications in language learning and communication systems. I am also investigating the synthesis of singing speech by modeling challenging singing signals like those of opera singers. From 2007 to 2015 I was Senior Researcher at the Telecommunications Research Center Vienna (FTW). Since 2016 I am Senior Research Scientist at the Acoustics Research Institute (ARI) of the Austrian Academy of Sciences (ÖAW).

Professional Experience

Since 2016	Senior Research Scientist at the Acoustics Research Institute (ARI) of the Austrian Academy of Sciences (ÖAW).
Since 2011	Lecturer at Vienna University of Technology.
2007 to 2015	Senior Researcher at Telecommunications Research Center Vienna (FTW).
2001 to 2007	Researcher at Telecommunications Research Center Vienna (FTW).
1999 to 2002	Software/database design and development with Java/Oracle.
1989 to 1993	Worked as a chef in restaurants in Austria and Liechtenstein.

Education

2015	Master degree (Dipl.-Ing.) in Computer Science from Vienna University of Technology.
2010 to 2015	Master's studies in Computer Science (Computational Intelligence) at Vienna University of Technology.
2007	Doctoral degree (Dr.techn.) in Electrical Engineering (with distinction) from Graz University of Technology.
2004 to 2007	Doctoral studies in Electrical Engineering (Speech Communication) at Graz University of Technology.
2001	Diploma degree (Mag.phil.) in Philosophy (with distinction) from the University of Vienna.
1995 to 2000	Diploma studies in Computer Science (Computational Logic) at Vienna University of Technology.
1994 to 2001	Diploma studies in Philosophy, Logic, and Mathematics at University of Vienna.
1984 to 1988	Cook apprenticeship.

Research visits

04 to 05 2014	National Institute of Informatics (NII), Tokyo, Japan.
08 to 09 2008	Centre for Speech Technology Research (CSTR), University of Edinburgh, UK.
08 2006	Telekom Innovation Laboratories (T-Labs), Berlin, Germany.
02 to 07 2005	International Computer Science Institute (ICSI), Berkeley, California.

Invited Talks

Invited talk on *Interpolation of language varieties in HMM-based speech synthesis*, 23. May 2014, NII SMG group, Tokyo, Japan.

Invited keynote talk on *Acoustic modeling, interpolation, and transformation of language varieties for speech synthesis* at International Dagstuhl Workshop on Multilinguality in Speech Research: Data, Methods and Models, 9.-11.4. 2014.

Publications

I have published more than 60 papers in international conferences and journals. The following are my most important publications. A full list can be found at http://sociolectix.org.

1. Cassia Valentini-Botinhao, Markus Toman, Michael Pucher, Dietmar Schabus, Junichi Yamagishi, Intelligibility of time-compressed synthetic speech: compression method and speaking style. *Speech Communication*, Volume 74, pp. 52-64, November 2015.
2. Markus Toman, Michael Pucher, Sylvia Moosmüller, Dietmar Schabus, Unsupervised interpolation of language varieties for speech synthesis. *Speech Communication*, Volume 72, pp. 176-193, September 2015.
3. Dietmar Schabus, Michael Pucher, Gregor Hofer, Joint Audiovisual Hidden Semi-Markov Model-based Speech Synthesis. *IEEE Journal of Selected Topics in Signal Processing*. Vol. 8, No. 2, pp. 336-347, April 2014.
4. Phillip L. De Leon, Michael Pucher, Junichi Yamagishi, Inma Hernaez, Ibon Saratxaga Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, Volume 20, Issue 8, October 2012, Pages 2280-2290.
5. Michael Pucher, Dietmar Schabus, Junichi Yamagishi, Friedrich Neubarth, Volker Strom, Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis. *Speech Communication*, Volume 52, Issue 2, February 2010, Pages 164-179.

Theses

- 2015, Michael Pucher, A Hidden-Markov-Model (HMM) based Opera Singing Synthesis System for German, Master thesis, Computer Science, Vienna University of Technology.
- 2007, Michael Pucher, Semantic Similarity in Automatic Speech Recognition for Meetings, Doctoral Thesis, Electrical and Information Engineering, Graz University of Technology.
- 2001, Michael Pucher, Formale Wahrheitstheorien nach Alfred Tarski, Diploma Thesis, Philosophy, University of Vienna.

Programming skills

C, C++, Java, SQL, Perl, Python, MATLAB, R, Shell scripting, Prolog, Lisp.

Projects

Research Projects

2013 to 2015 OPERA - Acoustic analysis and statistical modeling of Vienna opera singers, (as external collaborator) [NII - National Institute of Informatics, Japan - 118K].

2012 to 2016 AMTV - Acoustic modeling and transformation of varieties for speech synthesis (as principal investigator) [FWF: P23821-N23 - 296K].

2013 to 2015 SALB - Speech synthesis of auditory lecture books for blind children (as principal investigator) [BMWF - 116K].

2010 to 2014 AVDS - Adaptive Audio-Visual Dialect Synthesis (as principal investigator) [FWF: P22890-N23 - 299K].

2007 to 2009 VSDS - Viennese Sociolect and Dialect Synthesis (as principal investigator) [WWTF - 440K].

2009 to 2010 HI-MONI - Highway Monitoring (as project manager) [COMET].

2006 to 2007 TIDE - Testbed for Interactive Dialog System Evaluation (as project manager) [T-LABS].

Development Projects

2014 Bad Goisern and Innervillgraten Audio-Visual Dialect Speech Corpus (GIDS).

2014 Release of SALB – a frontend for speech synthesis using HTS voice models

2013 Release of Austrian German open source HTS voice.

2010 Development of "Leopold" the first synthetic voice for Austrian German together with company partners, which was integrated into a web reading service for the Website of the City of Vienna.

International Scientific Cooperation Partners

- Prof. Junichi Yamagishi – University of Edinburgh, UK / National Institute of Informatics, Japan.
- Prof. Inma Hernaez – University of the Basque Country, Bilbao, Spain.
- Prof. Phillip de Leon – New Mexico State University, USA.

Teaching

SS 2013 to 2016	Lecture on Cognitive User Interfaces at Institute of Computer Languages at Vienna University of Technology.
WS 2011 to 2013	Lecture on Computational Semantics at Institute of Computer Languages at Vienna University of Technology.
SS 2011	Seminar on *Audio-Visual Speech Synthesis* at the Signal Processing Laboratory (Aholab) of the University of the Basque Country.
SS 2008	Seminar on *Speech Synthesis* at the Signal Processing and Speech Communication Laboratory (SPSC Lab) at Graz University of Technology.

I co-supervised the following PhD theses at FTW:

- 2016, Markus Toman, Acoustic modeling and transformation of varieties for speech synthesis (Vienna University of Technology).
- 2014, Dietmar Schabus, Adaptive audio-visual speech synthesis (Graz University of Technology).

I co-supervised the following diploma theses at FTW

- 2013, Jakob Hollenstein, Visual Control of Hidden-Semi-Markov-Model based Acoustic Speech Synthesis, Diploma thesis. Vienna University of Technology, 2013.
- 2009, Dietmar Schabus, Interpolation of Austrian German and Viennese Dialect / Sociolect in HMM-based Speech Synthesis. Diploma thesis. Vienna University of Technology, 2009.
- 2008, Christian Kranzler, Text-to-Speech Engine with Austrian German corpus. Diploma thesis. Graz University of Technology, 2008.

- 2008, Michael Bruss, Quantitative und phonetische Analyse von nicht-linguistischen Partikeln in spontan gesprochener Sprache der Wiener Soziolekte. Magisterarbeit. Universität des Saarlandes, Saarbrücken, 2008.

Professional Activities

Organizing

- Area chair for *Speech Synthesis and Spoken Language Generation* of INTERSPEECH 2015.
- Organizing committee member of FAAVSP 2015 - The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing.
- Program committee member of ACM Multimedia 2014.
- Organizing committee member of FAA 2012 - The ACM 3rd International Symposium on Facial Analysis and Animation.

Reviewing

- Speech Communication, Elsevier
- IEEE Transactions on Audio, Speech, and Language Processing
- Computer Speech and Language, Elsevier
- IEEE Signal Processing Letters
- IEEE Transactions on Systems, Man, and Cybernetics
- Cognitive Computation, Springer

Memberships

- IEEE
- ACM
- International Speech Communication Association (ISCA)
- European Network for the Advancement of Artificial Cognitive Systems, Interaction and Robotics (EUCOG III)
- Cross-Modal Analysis of Verbal and Nonverbal Communication (COST 2102)