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

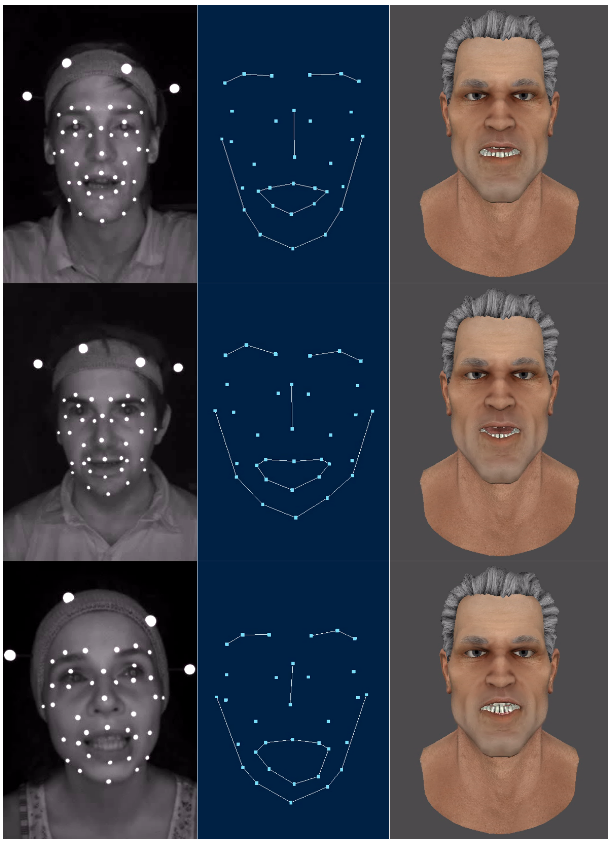

Figure 1: *Still images from the recording session (left), the corresponding 3D marker data (middle) and the resulting pose of the virtual head with this data applied (right). See also videos at http://userver.ftw.at/~schabus/interspeech2012/*

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/

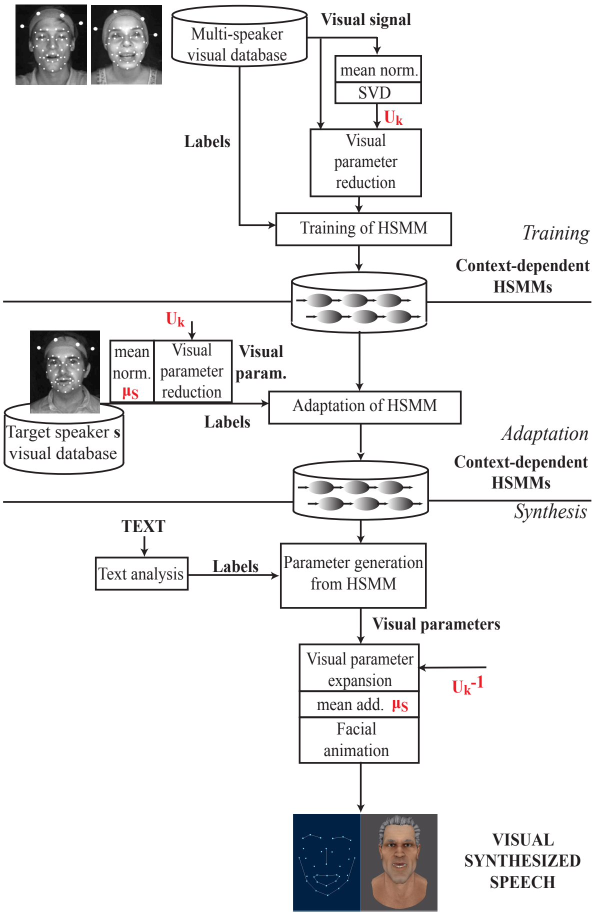

Figure 2: *Overview of the speaker-adaptive visual modeling framework.*

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.

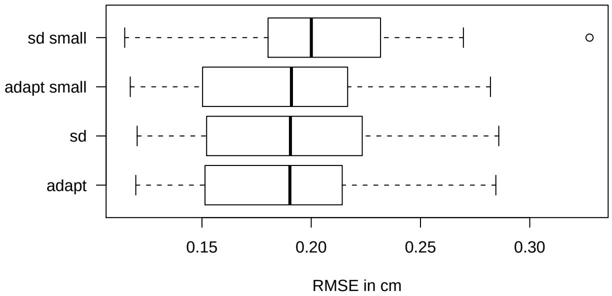

Figure 3: *Box plots of the root mean squared differences between synthesized and recorded marker positions.*

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.

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.

[2] See example stimuli at http://userver.ftw.at/~schabus/interspeech2012/