Building a Synchronous Corpus of Acoustic and 3D Facial Marker Data for Adaptive Audio-visual Speech Synthesis

Dietmar Schabus[1,2], Michael Pucher[1], Gregor Hofer[1]

[1]FTW Telecommunications Research Center Vienna
Donau-City-Strasse 1, 1220 Vienna, Austria
schabus@ftw.at, pucher@ftw.at, hofer@ftw.at

[2]SPSC Lab, Graz University of Technology
Graz, Austria

Abstract

We have created a synchronous corpus of acoustic and 3D facial marker data from multiple speakers for adaptive audio-visual text-to-speech synthesis. The corpus contains data from one female and two male speakers and amounts to 223 Austrian German sentences each. In this paper, we first describe the recording process, using professional audio equipment and a marker-based 3D facial motion capturing system for the audio-visual recordings. We then turn to post-processing, which incorporates forced alignment, principal component analysis (PCA) on the visual data, and some manual checking and corrections. Finally, we describe the resulting corpus, which will be released under a research license at the end of our project. We show that the standard PCA based feature extraction approach also works on a multi-speaker database in the adaptation scenario, where there is no data from the target speaker available in the PCA step.

Keywords: speech synthesis, audio-visual, corpus

1. Introduction

Audio-visual TTS is the synthesis of both an acoustic speech signal (TTS in the classical sense), as well as a matching animation sequence of a talking face, given some unseen text as input. Since we target animation synthesis in 3D, unlike video-based photo-realistic methods, we need to capture 3D information during recording.

Although our goal is primarily audio-visual TTS, a synchronous multi-modal corpus, as well as the the process of building it, can be of relevance to other fields as well (e.g., audio-visual speech recognition, synthesis of blinking, eye brow movement or head motion, multi-modal recognition of speaker, language variety or emotion, etc.).

This corpus will be used primarily as training data in the statistical parametric framework (perhaps better known as HMM-based speech synthesis) (Tokuda et al., 2008; Zen et al., 2009), where both acoustic speech parameters and animation parameters can be generated by a maximum likelihood parameter generation algorithm (Tokuda et al., 2000). With a small test corpus, we have already demonstrated the feasibility of such an approach (Schabus et al., 2011). As for any data-driven method, the aim here is to create a sufficiently large corpus of high quality while keeping the required time and effort as small as possible. The corpus contains multiple speakers because we are also investigating average voices and speaker adaptation (Yamagishi et al., 2009).

Audio-visual corpora described in the literature often consist of simultaneous audio and single-view video camera recordings (e.g., (Hazen et al., 2004; Cooke et al., 2006; Theobald et al., 2008)). While such corpora are an important asset for the research community, they are unsuitable for synthesis in 3D, which is the focus of this work.

Visual synthesis in 3D has been investigated intensively, also for German language (Fagel and Bailly, 2008) and also based on HMMs (Govokhina et al., 2007). In contrast to these other works, our focus is on the adaptive scenario and hence we require a multi-speaker corpus.

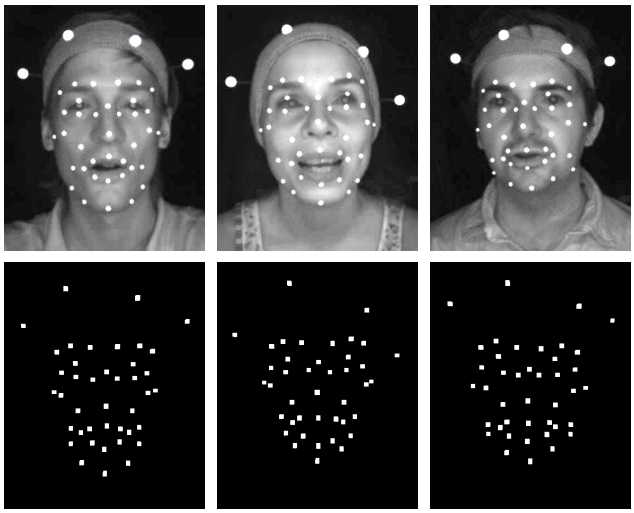

Figure 1: Still images from grayscale videos showing facial marker layout (top) for 3 different speakers and corresponding renderings of 3D marker data (bottom).

2. Recordings

We have recorded three 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 some prosodic variance (phrase breaks etc.). It amounts to 223 utterances and roughly 11 minutes total for each of the speakers. For acoustic synthe-

sis, we would need to use an adaptive approach that combines data from multiple speakers and varieties in the background model, to produce high-quality voices with such small corpora.

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 96 kHz sampling rate, 24 bit encoding, and a professional microphone (an AKG C-414 B-TL). The acoustic recordings were later downsampled to 44.1 kHz and 16 bit encoding. 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 HMM-based speech synthesis (Yamagishi and King, 2010).

For the recording of facial motion, we used a commercially available system called OptiTrack[1]. 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 (top) for still images from the grayscale video showing the marker layout.

For synchronization between the audio and 3D recordings, we use a simple clapping signal at the beginning of each take. This makes it straightforward to identify the position of the signal in both the audio recordings as well as in the grayscale video. Due to the frame rate of the latter, we accomplish a synchronization accuracy of ±5 ms. Each recording session was started with a neutral pose (relaxed face, mouth closed, eyes open, looking straight ahead).

The OptiTrack software stores the recorded data using its own format, but it can export to the open C3D format as well as to the proprietary but widespread FBX format. To ease post-processing (see next section), we have chosen to convert and store the 3D data in a more simplistic format. Each of the 41 markers has a name (e.g., *LMouthCorner*) and a (x, y, z) position for each recorded frame. We stack all the coordinates vertically, in alphabetical marker name order, to form a column vector of $41 \cdot 3 = 123$ entries which describes a single frame. We then stack such frame column vectors horizontally, forming matrices of shape $123 \times n$, where n is the number of frames (and $n/100$ is the duration of the utterance in seconds). Hence, each row of the matrix gives the trajectory of a certain coordinate of a certain marker over time.

3. Adaptive Audio-Visual Modeling

We record a multi-speaker audio-visual database to perform adaptive audio-visual modeling. Multiple speakers are used to train an average audio-visual model using speaker adaptive training (SAT) (Figure 2). At adaptation time, audio-visual data of a certain speaker is used to adapt the average model. At synthesis time, we generate a synchronized acoustic and visual sequence. The advantage of the adaptive approach is the possibility to use an average (background) model that is trained on a large amount of training data, hence requiring only a small amount of adaptation data from the target speaker.

[1] http://www.naturalpoint.com/optitrack/

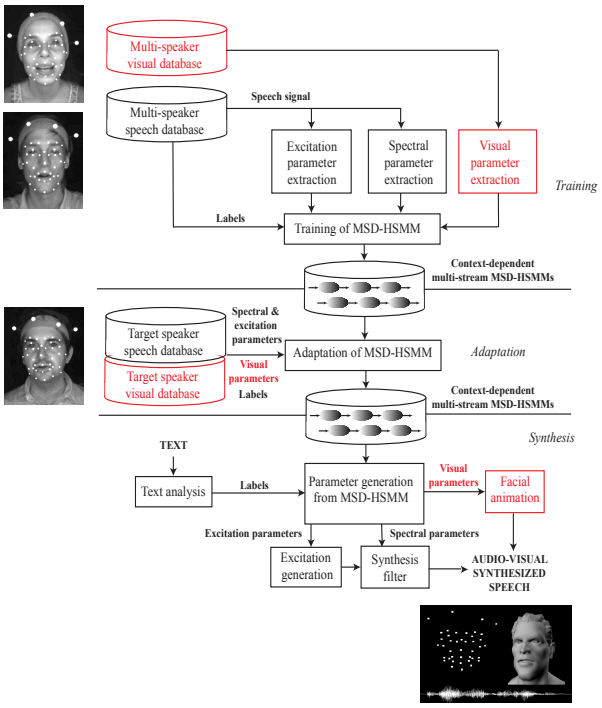

Figure 2: Adaptive HMM-based audio-visual speech synthesis.

4. Post-processing

Since the 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. 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. After this step, the four headband markers become static and can be removed, leaving 37 markers. Furthermore, we have removed the four markers on the upper and lower eyelids, since we believe that phones as modeling units are inappropriate for modeling eye blinking, i.e., we would use a separate model with different modeling units for synthesizing eye blinks. This leaves 33 markers, and hence 99-dimensional frame representations.

If we choose a reference frame with a neutral pose (i.e., one of the first frames, see above) for each recording session, this reference can also be used for positional normalization. We have decided to translate all recording sessions, such that the position of a certain marker (central upper lip) is the same for the neutral poses across all sessions and speakers.

The acoustic recordings were cut into single utterances semi-automatically and then annotated on the phone level using hidden-Markov-model based flat-start forced alignment with mono-phone models using HTK (Young et al., 2006). The resulting alignment was checked by looking at the phone borders in the spectrogram of each utterance

and the few obvious mistakes were corrected manually. Furthermore, we have trained standard (audio) speaker-dependent models using HTS, and the synthesis results confirm that the alignment is adequate.

Using the utterance border information from the audio data together with the synchronization offsets from the clapping signal, we extract the corresponding frames for each utterance from the 3D recordings as well as from the grayscale videos.

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 our 99-dimensional vectors allow. Guided by this intuition, as well as to de-correlate the components, we have carried out principal component analysis (PCA) on our 3D data, in a manner similar to the well-known eigenface approach (Turk and Pentland, 1991a; Turk and Pentland, 1991b). For a single speaker, we can look at the matrix M of size $9 \times n$ of all frames of all utterances of that speaker stacked horizontally, subtract the sample mean column vector μ from each column of M to obtain a normalized $\bar{M}$, and compute the singular value decomposition (SVD):

$$\bar{M} = U \cdot \Sigma \cdot V^T$$

We are solely interested in the matrix U of size 99×99, whose columns are the bases of the principal component space, sorted by decreasing eigenvalues.

In order to determine how many bases (dimensions) are needed to adequately represent the data, we can calculate the reconstruction error when we use only the first k principal components ($k \in [1, 99]$). Let U_k denote the matrix containing only the first k columns of a matrix U from an SVD of a matrix M. Then we define the reconstruction of a matrix N of size $99 \times n$ as

$$\bar{N}_{rec} = U_k \cdot U_k^T \cdot \bar{N}$$

Re-adding N's sample mean gives us N_{rec}, and we can compute the error matrix $E = N - N_{rec}$. Finally, we define the reconstruction error as the root mean squared error (RMSE) across all elements e_{ij} of E:

$$\text{RMSE} = \sqrt{\frac{1}{99n} \sum_{i=1}^{99} \sum_{j=1}^{n} {e_{ij}}^2}$$

Note that the matrix M used to compute U via SVD and the reconstructed matrix N can be the same or different. In fact, we are especially interested in cases where they are not the same. Keeping in mind that we are interested in training average voices and adapting these to target speakers, we have calculated the RMSE for the following scenarios: We always consider one of our three speakers (*dsc*, *mpu* and *nke*) as the target speaker, i.e., the data to be reconstructed are all frames of all utterances of that speaker. The data used to determine the transformation into principal component space (matrix U) is either

1. the data from the target speaker
2. the data from all three speakers (including the target speaker)
3. the data from the two other speakers (excluding the target speaker).

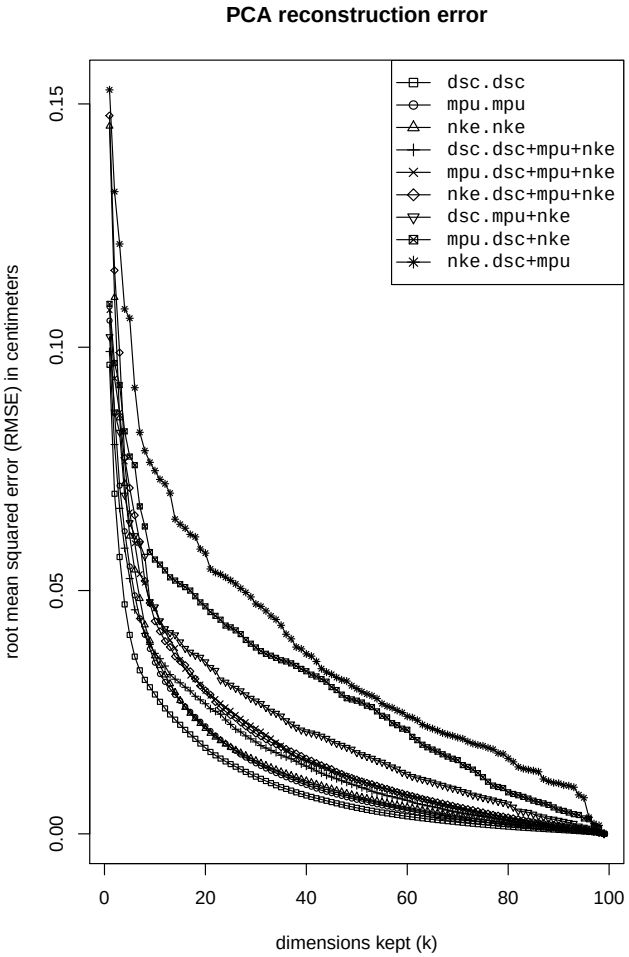

Figure 3: PCA reconstruction error (RMSE) for the nine different conditions and varying k.

Especially the third case is of high relevance in an adaptation scenario, as the data of the target speaker is typically not part of the data for the average voice.

Fig. 3 shows the RMSE reconstruction error for each of these nine conditions and for all $k \in [1, 99]$. The points are labeled with the target speaker before the period and all speakers that were used in the SVD after the period. Overall, we see our intuition confirmed: using only 6 of 99 dimensions yields an RMSE of less than 1 mm in all nine conditions. The three speaker-specific versions produce the best results, as expected. Their RMSEs lie even below 0.6 mm at $k = 6$ and below 0.25 mm at $k = 18$. The three versions with all speakers in the SVD are a bit worse than that, and as expected the three held-out versions yield the worst results. It takes 59 dimensions for the particularly bad *nke.dsc+mpu* to reach an RMSE below 0.25 mm.

In general, we have the positive result that it is possible to project some speaker's data into a much smaller subspace, where the definition of the subspace and the projection into it were determined without using any data from that speaker, without making a large reconstruction error.

We would also expect the results to improve once we have data from a larger number of speakers available.

5. Corpus Description

The final corpus consists of 223 utterances read by one female and two male speakers. The sentences are from the well-known Kiel Corpus of Read Speech (IPDS, 1994; Brinckmann, 2004) and include 100 "Berlin", 100 "Marburg", 16 "Buttergeschichte" and 7 "Nordwind und Sonne" sentences. This amounts to roughly 11 minutes per speaker. For each utterance, the corpus contains:

- A Wave audio file (44.1 kHz, 16 bit encoding)
- A Python pickle file[2] containing the 3D-coordinates of each facial marker at each frame ($41 \cdot 3$ dimensions, 100 frames per second)
- A Python pickle file containing the same data at reduced dimensionality after PCA (30 dimensions, 100 fps)
- An AVI video file containing grayscale footage as shown in the top part of Fig. 1 (640×480 pixels, 100 fps XviD, Wave audio)
- An HTK mono-phone label file, providing the transcription and precise temporal phone borders
- An HTK full-context quin-phone label file, additionally providing phonetic context and sentence structure information

Furthermore, the corpus contains the transformation matrices from PCA used in the nine conditions described in the previous section, as well as a Python program that can play back the 3D data graphically (see bottom part of Fig. 1).

6. Summary

We have shown how to develop an audio-visual multi-speaker corpus for adaptive audio-visual speech synthesis. Our feature extraction results show that we can use the standard PCA approach for feature extraction in the visual modality in an adaptation setting.
In future work we will extend this corpus to Austrian dialect varieties from the middle-Bavarian and south-Bavarian dialect regions. At the end of our project[3], we plan to release the corpus under a research license.
Using this data, we want to explore adaptive visual and audiovisual speech synthesis in the HTS framework.

7. Acknowledgements

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.

[2] http://docs.python.org/library/pickle.html
[3] https://portal.ftw.at/projects/avds

8. References

C. Brinckmann. 2004. The Kiel corpus of read speech as a resource for speech synthesis. Master's thesis, Universität des Saarlandes, Saarbrücken, Germany.

M. Cooke, J. Barker, S. Cunningham, and X. Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424.

S. Fagel and G. Bailly. 2008. German text-to-audiovisual-speech by 3-d speaker cloning. In *International Conference on Auditory-Visual Speech Processing (AVSP)*, Tangalooma, QLD, Australia.

O. Govokhina, G. Bailly, and G. Breton. 2007. Learning optimal audiovisual phasing for a HMM-based control model for facial animation. In *6th ISCA Workshop on Speech Synthesis (SSW6)*, pages 1–4, Bonn, Germany.

T. J. Hazen, K. Saenko, C. La, and J. R. Glass. 2004. A segment-based audio-visual speech recognizer: data collection, development, and initial experiments. In *Proc. ICMI*, pages 235–242, State College, PA, USA.

IPDS. 1994. The Kiel corpus of read speech. University of Kiel, Germany, http://www.ipds.uni-kiel.de/publikationen/kcrsp.en.html.

D. Schabus, M. Pucher, and G. Hofer. 2011. Simultaneous speech and animation synthesis. In *ACM SIGGRAPH 2011 Posters*, Vancouver, BC, Canada.

B. Theobald, S. Fagel, G. Bailly, and F. Elisei. 2008. Lips2008: visual speech synthesis challenge. In *Proc. Interspeech*, pages 2310–2313, Brisbane, QLD, Australia.

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

K. Tokuda, H. Zen, J. Yamagishi, A. Black, T. Masuko, S. Sako, T. Toda, T. Nose, and K. Oura. 2008. The HMM-based speech synthesis system (HTS).

M. Turk and A. Pentland. 1991a. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71 – 86.

M. Turk and A. Pentland. 1991b. Face recognition using eigenfaces. In *Proc. CVPR*, pages 586 –591, Maui, HI, USA.

J. Yamagishi and S. King. 2010. Simple methods for improving speaker-similarity of HMM-based speech synthesis. In *Proc. ICASSP*, pages 4610–4613, Dallas, TX, USA.

J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, and S. Renals. 2009. Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 17(6):1208–1230.

S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. 2006. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK.

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