Objective and Subjective Feature Evaluation for Speaker-Adaptive Visual Speech Synthesis

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

[1]Telecommunications Research Center Vienna (FTW), Vienna, Austria
[2]Graz University of Technology, Graz, Austria
{schabus,pucher,hofer}@ftw.at

Abstract

This paper describes an evaluation of a feature extraction method for visual speech synthesis that is suitable for speaker-adaptive training of a Hidden Semi-Markov Model (HSMM)-based visual speech synthesizer. An audio-visual corpus from three speakers was recorded. While the features used for the auditory modality are well understood, we propose to use a standard Principal Component Analysis (PCA) approach to extract suitable features for training and synthesis of the visual modality. A PCA-based approach provides dimensionality reduction and component de-correlation on the 3D facial marker data which was recorded using a facial motion capturing system. Enabling visual average "voice" training and speaker-adaptation brings a key strength of the HMM framework into both the visual and the audio-visual domain. An objective evaluation based on reconstruction error calculations, as well as a perceptual evaluation with 40 test subjects, show that PCA is well suited for feature extraction from multiple speakers, even in a challenging adaptation scenario where no data from the target speaker is available during PCA.

1. Introduction

The features used to parametrize the acoustic speech signal for such a training-synthesis pipeline are fairly well established. In this paper, we aim to justify the use of standard PCA features as suitable for the visual domain. While doing so, we pay special attention to the feasibility of adaptive training across multiple speakers. The field of visual speech synthesis is also 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, 5] have been developed. While especially video-based systems produce quite convincing animation, they require large amounts of training data. Furthermore none of these approaches facilitate adaption of models between different speakers like it is possible for speech synthesis using the HMM framework. The HMM-based visual speech synthesis systems that have been developed can be broadly categorized into two types: image-based systems or motion capture based systems. Image-based systems use features derived directly from the video frames [6] where the resulting synthesis is supposed to look like a video of a real person. These types of features usually require the synthesized trajectories to be played back on the same face that was recorded, which makes them unsuitable for multi-speaker adaption. Motion capture based approaches [7, 8] derive their features from individual 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, which makes them highly suitable for adaptation. However, to the best of our knowledge the adaptation of speech data in the visual domain has not been investigated, except in our own work [9], where the results now reported in this paper where in fact already used.

Speaker-adaptive visual and joint audio-visual speech synthesis can be applied in some video games, which can have hundreds of different characters, each speaking only a few lines of dialogue. Other applications for this type of technology include multi-modal communication or any other application which requires many different speakers and facial models.

2. Description of Data and Usage Scenario

We have recorded three speakers reading the same recording script in standard Austrian German to create a synchronized audio-visual corpus. It amounts to 223 utterances and roughly 11 minutes total for each of the speakers. For the recording of facial motion, we used a commercially available system called OptiTrack [10] which records the 3D position of 37 reflective markers glued to a person's face at 100 Hz. Figure 1 shows the marker layout and the resulting 3D data for an example frame of each of the three speakers. After subtracting rigid head motion and removal of the upper and lower eyelid markers, we have 99-dimensional (33 markers $\times$ 3 spatial dimensions) face representations changing over time. Using this kind of data (recorded or synthesized), the movement of a virtual head can be controlled by means of marker retargeting to a facial rig, a common

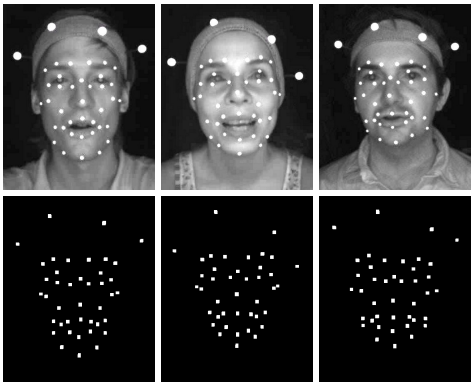

Figure 1: Still images from grayscale videos showing facial marker layout (top) for 3 different speakers and corresponding renderings of 3D marker data (bottom).

technique in animation. We have published a more detailed description of this corpus before [11].

We plan to use this data for speaker-adaptive training (SAT) [12] based on constrained maximum likelihood linear regression (CMLLR) [13] in an HSMM-based speech synthesis framework [14] for visual and audio-visual speech synthesis. For acoustic synthesis it was shown that this method can produce voices that are similar to a target speaker by using only a small amount of adaptation data [15]. To employ this method in the visual domain, we need to show that the visual feature extraction method is usable in the adaptation setting.

Figure 2 illustrates a speaker-adaptive audio-visual speech synthesis system. 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.

3. PCA-based Feature Extraction

We have already described PCA-based feature extraction in detail in our previous paper on the data corpus [11]. To briefly summarize, the idea is to carry out a projection of the visual data into a lower-dimensional PCA-space (making a small reconstruction error) and at the same time to de-correlate the components. To do so, we first subtract the sample mean column vector μ_s from the data matrix M_s to obtain a mean-normalized $\bar{M}_s$ for each speaker $s \in AVG = \{s_1, s_2, \ldots\}$. Then all speakers' normalized data is combined to one big matrix $\bar{M}_{AVG}$, on which we compute the singular value decomposition (SVD):

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

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. We can project a frame column vector x from $\bar{M}_{AVG}$ into principal component space and back using U:

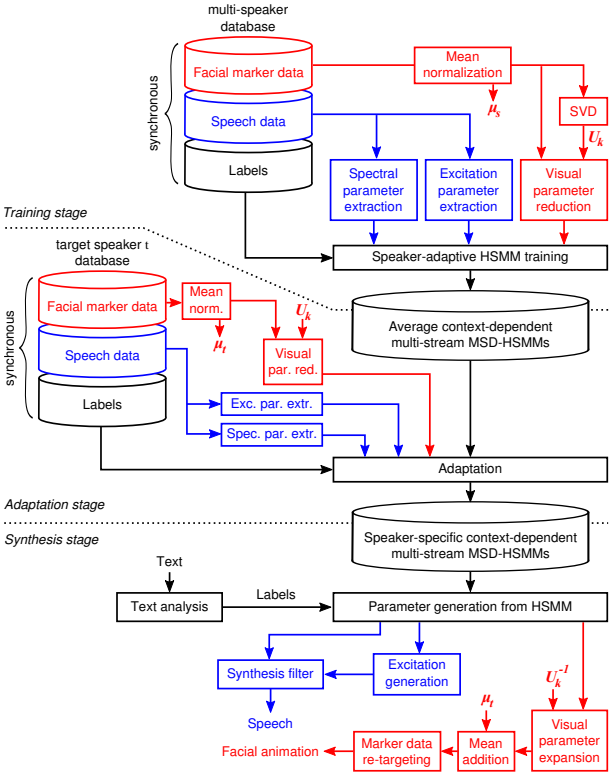

Figure 2: Overview of an adaptive audio-visual speech synthesis system, which consists of four main components: audio-visual speech analysis, average audio-visual training, speaker adaptation, and audio-visual speech generation.

$$x = U \cdot (U^{-1} \cdot x)$$

Furthermore, if U_k denotes the matrix containing only the first k columns of U, then we can approximate this equality:

$$x \approx U_k \cdot (U_k^{-1} \cdot x),$$

which provides a projection function to and from a k-dimensional subspace ($k < 99$) in which HSMM training, adaptation and synthesis can be carried out.

The key idea regarding the speaker-adaptive scenario is now to apply this same projection, which was determined on the data for the average voice, to project the adaptation data from the target speaker t into the same subspace. This assumes that we find a subspace via SVD on the data from the (potentially large number of) speakers in the average voice that is general enough to also contain the target speaker's data, provided that we do not choose k too "tight". The purpose of this paper is to justify this assumption, as well as to choose an appropriate value for k.

Specifically, for the data we have recorded, we always consider one of our three speakers (dsc, mpu and nke)

as the target speaker, i.e., the data to be projected (and reconstructed, when we talk about the reconstruction error later on) are all frames of all utterances of that speaker. The data used for SVD, i.e., for calculating the projection into principal component space 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).

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 training data for the average voice. Intuitively, we expect this to be the most challenging of the three scenarios. But also the second case can be of practical relevance: when we want to put all available data to optimal use, it might be beneficial to include the target speaker in the average voice.

4. Evaluation

In order to evaluate how well the results of PCA, when carried out the way we have described in the previous section, do match our task of visual feature extraction, we use both objective and subjective performance measures. The following subsection considers the objective reconstruction error. This should bring insight to the behavior of the three methods mentioned in the last section, to understanding the role of certain markers, as well as to the influence of k, the number of kept dimensions. The subsection after that presents the results of a subjective evaluation we have carried out with 40 test subjects. The main purpose of this is to provide a basis for deciding on the value of k. We then discuss and compare the two measures in a third subsection.

4.1. Objective Evaluation via Reconstruction Error

Given a matrix U_k containing the first k columns of a matrix U resulting from SVD (as described in Section 3), we define the reconstruction of a data matrix M, containing a target speaker's utterances stacked horizontally, as

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

Re-adding M's sample mean to $\bar{M}_{rec}$ gives us M_{rec}, and we can compute the error matrix $E = M - M_{rec}$. Let n denote the total number of frames in all utterances of the target speaker, i.e., M, $\bar{M}$, $\bar{M}_{rec}$, M_{rec} and E are all of size $99 \times n$, while U_k is of size $99 \times k$. Finally, we define the reconstruction error as the root mean squared error (RMSE) across all elements e_{ij} of E:

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

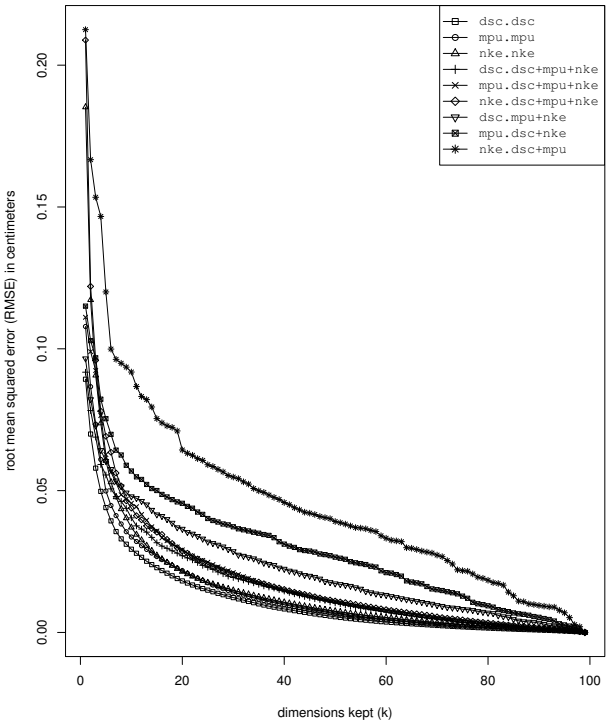

Figure 3: PCA reconstruction error (RMSE) for the nine different conditions and varying k.

We have computed the RMSE for all $k \in \{1, \ldots, 99\}$ and for each of the nine conditions resulting from the combination of each of our three speakers as target speaker with one of the three methods to compute the SVD as described in Section 3. The results are shown in Figure 3. 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 1mm in all nine conditions. The three speaker-specific versions produce the best results, as expected. Their RMSEs lie even below 0.5mm at $k = 6$. 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 35 dimensions for the particularly bad *nke.dsc+mpu* to reach an RMSE below 0.5mm.

Although the methods of the third kind produce a larger reconstruction error than the others, they still show the same overall behavior (shape of the curves in Fig. 3), namely that the first few dimensions make a very big difference in the results, and that the error levels off towards the larger values of k. This means that 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, given that we do not choose

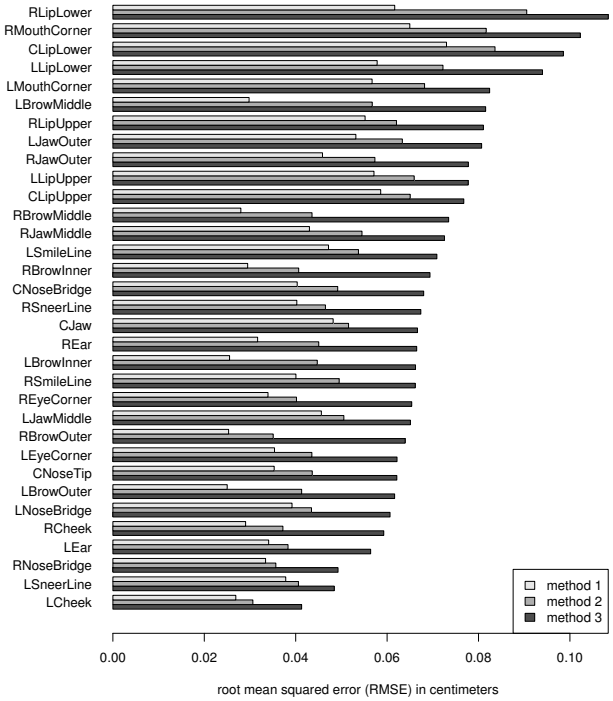

Figure 4: PCA reconstruction error (RMSE) for each marker and SVD method, averaged across all target speakers and 3D coordinates, at $k = 6$.

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

Rather than taking the mean across the entire error matrix E, we can also look at the means of each row, which corresponds to the mean error for a certain coordinate of a certain marker. Fig. 4 shows the RMSE for each marker and each of the three methods. The plotted values are means across all target speakers, 3D coordinates and of course frames, for a fixed value of $k = 6$. We can see that the markers in the region of the mouth (**Lip**, **Mouth**, **Jaw**) are responsible for the largest errors. Also, we see again how the third method (held-out) is consistently worse than the second (all speakers), which is in turn consistently worse than the first (speaker-specific).

4.2. Subjective Evaluation via Perceptive Experiments

Based on the objective evaluation alone, it would be difficult to choose a value for k to proceed to actual training and synthesis. It is not clear a priori what an RMSE of, e.g., 1mm means perceptually, or in other words it is not clear how small we can choose k without perceived degradation in quality. To clarify this, we have carried out a subjective perceptual experiment with 40 non-expert test subjects (half females and half males, aged 20–68 years). This experiment was designed as follows.

Table 1: Partitioning of the values for k for target speaker *nke*

bin	method 1	method 2	method 3
1	1	1	1
2	2	2	2–3
3	3	3	4–5
4	4	4–5	6–8
5	5–6	6–7	9–11
6	7–8	8–9	12–14
7	9–10	10–11	15–17
8	11–12	12–14	18–21
9	13–15	15–17	22–25
10	16–18	18–20	26–29
11	19–21	21–24	30–33
12	22–25	25–28	34–38
13	26–29	29–33	39–43
14	30–34	34–38	44–49
15	35–40	39–44	50–55
16	41–47	45–52	56–62
17	48–56	53–62	63–70
18	57–70	63–78	71–80
19	71–99	79–99	81–99
20	99	99	99

We have created videos of marker renderings, where for each frame of an utterance, a white cube is drawn on a black background for each of the 33 markers at the 3D position of that marker in that frame[1]. This leads to renderings that look similar to the lower part of Fig. 1. Note that we deliberately chose not to apply the marker motion to a virtual head and use renderings of the animated head in the evaluation, because we wanted to make sure the quality (or lack of quality) of the retargeting or the visual appearance of the head do not skew the evaluation results.

In each video, we showed a rendering of the originally recorded data side by side with a rendering of a reconstruction using a certain value of k. Then the test subjects were asked to decide whether the two renderings were *different* or *the same* from their point of view. Whether the original was on the left or on the right was chosen randomly for each video. We used the first five sentences of our corpus as test sentences, and each test subject saw one comparison for each test sentence and each of the nine conditions (Fig. 3), i.e., 45 comparisons in total. We have selected values of k with respect to the reconstruction error: For each of the nine conditions, we have partitioned the set of 99 possible values for k into 19 bins, where each bin amounts for a similar percentage of the overall error. We also added a twentieth bin containing only the last value ($k = 99$). We then selected the middle value of each bin as that bin's representative. Each test subject saw at least one comparison from each bin, with the remaining comparisons distributed randomly. Table 1 shows for target speaker *nke* which values of k belonged to which of the 20 bins in each of the three methods.

[1] Examples on http://userver.ftw.at/~schabus/avsp2013fe/examples.mov

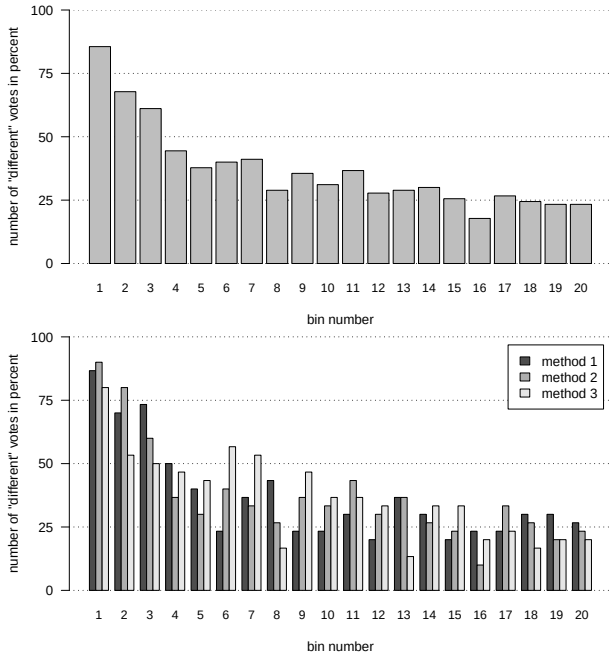

Figure 5: Results of the subjective evaluation: Percentage of "different" votes per bin (top) and per bin and method (bottom).

This leads to a denser sampling in the lower region, where one additional dimension makes a big difference, and a sparser sampling in the higher region, where one additional dimension makes a small difference. The entire evaluation thus amounts to 900 comparisons (9 methods $\times$ 5 sentences $\times$ 20 bins), for each of which we have two votes from two different subjects. Therefore the results contain 1800 votes total.

The results are shown in Fig. 5, where we have plotted the percentage of "different" votes for each of the 20 bins (top), and the same data additionally separated by method (bottom). We see that the general picture is in agreement with what we know from the objective error, namely that reconstructions with low values of k (left side of Fig. 5) are perceived as being mostly different from the original, that small changes to k have a shrinking influence with growing k, and that the difference levels off towards the upper end of the scale (right side of Fig. 5).

However, the actual values of the evaluation at the extreme points are somewhat surprising: The reconstructions corresponding to the first bin are very poor in terms of the objective error and should look clearly different from the original, yet in 13 of the 90 comparisons (14%) they were perceived as being equal by the test subjects. Similarly, at the other end of the scale, the reconstructions with $k = 99$ in bin 20 are per definition error-free, as the projection into principal component space and back are mere rotations of the coordinate system.[2] Nevertheless, in 21 out of 90 cases (23%) they were judged as being different from the original.

[2]The actual RMSE in our implementation was always $< 10^{-15}$

Table 2: Significant differences in perception: Results of paired Wilcoxon signed rank tests between votes for each bin, with (■) and without Bonferroni correction (□).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
2	□			□	■	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■
3	■			□	□	□	□	■	□	■	□	■	■	■	■	■	■	■	■	■
4	■	□	□					□				□	□	□	□	■	□	□	□	□
5	■	■	□													□		□	□	□
6	■	□	□												□	□		□	□	□
7	■	□	□												□	□	□	□	□	□
8	■	■	■	□																
9	■	■	□													□				
10	■	■	■													□				
11	■	■	□													□		□	□	□
12	■	■	■	□																
13	■	■	■	□																
14	■	■	■	□																
15	■	■	■	□		□	□													
16	■	■	■	■	□	□	□		□	□	□									
17	■	■	■	□			□													
18	■	■	■	□	□	□	□				□									
19	■	■	■	□	□	□	□				□									
20	■	■	■	□	□	□	□				□									

We believe some of this uncertainty in the results can be ascribed to the difficulty of the task. Even if the marker motion is quite different from the original for low values of k, the overall appearance of the two renderings is very similar. Furthermore, the sequence of comparison examples is quite uniform, which could lead to effects of boredom.

This uncertainty also makes it difficult to compare the three methods to each other based on the subjective data. The bottom part of Fig. 5 illustrates that the data does not allow for drawing clear conclusions in this regard.

To assess the statistical significance of the differences between the bins' results, we have computed Bonferroni-corrected paired Wilcoxon signed rank tests between the votes of each pair of bins. The pairing of votes was based on the method and utterance only, i.e., we ignored which test subject cast a particular vote. The results are shown in Table 2, where the symbol ■ indicates a significant difference ($\alpha = 0.05$). In this rather restrictive setting (due to Bonferroni correction the value of α for each of the 190 tests is $0.05/190 \approx 0.00026$), only the first four bins show significant differences from some of the other bins, i.e., none of the bins from 5 to 20 differ significantly from each other.

This result tells us that we need to choose k from a bin ≥ 4 at the very least, and it even suggests that choosing from bin number 4 is sufficient, since larger values do not lead to significantly better results anyway. However, the conservativeness of Bonferroni correction would act in our advantage here, because it reduces the probability of false positives (type I error) at the cost of an increased probability of false negatives (type II error). We should not choose k too small because of some signifi-

cant differences that were missed due to the Bonferroni correction. Therefore, Table 2 also shows the additional significances of the same test without Bonferroni correction, indicated by the symbol □. This result is quite likely to contain some false positives, but there is nevertheless the set of bins $\{12, ..., 20\}$ where there are no significant differences. Therefore, by selecting the smallest k larger than any k from bin 11 we still make a conservative choice. However, the final $k = 33$ still accounts for a great reduction in dimensionality: Two thirds of the initial 99 degrees of freedom could be removed.

4.3. Discussion

Overall, both the objective and the subjective evaluation have provided results in general agreement with the expectations. With growing k, the results improve quickly at first, and finally level off – towards zero in the objective case and towards "background noise" of uncertainty in the subjective case. The reconstruction error evaluation clearly showed the difference in performance between the three methods, something which the subjective method failed to show. However, the user votes provide an excellent basis for selecting an actual value for k that defines the number of dimensions employed in both training and synthesis.

5. Conclusion

We have shown that a PCA-based feature extraction algorithm in the visual domain is suitable for speaker-adaptive training. Overall we can conclude that it is feasible 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. This opens up the possibility of not only using speaker-adaptive training in the auditory domain but also extend it to the visual and joint audio-visual domains. With this approach we are able to adapt an average visual model to a specific speaker by using only a small amount of visual adaptation data, cutting down the time and effort required to produce a speech motion model for that new speaker.

For visual-only speech synthesis, we have already shown that the adaptive approach can be superior to speaker-dependent modeling, at least for small amounts of target speaker data [9]. We plan to extend this to joint audio-visual speaker-adaptive modeling in the future. Furthermore, we have worked on language varieties (dialects/sociolects) in the acoustic domain in the past, and we would like to extend some of these results to the visual modality.

6. Acknowledgments

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] M. Cohen and D. Massaro, "Modeling coarticulation in synthetic visual speech," *Models and Techniques in Computer Animation*, Jan 1993.

[2] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: driving visual speech with audio," in *Proc. SIGGRAPH '97*, Los Angeles, CA, USA, 1997, pp. 353–360.

[3] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *Proc. SIGGRAPH '02*, San Antonio, TX, USA, 2002, pp. 388–398.

[4] G. Bailly, M. Bérar, F. Elisei, and M. Odisio, "Audiovisual speech synthesis," *International Journal of Speech Technology*, 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. SIGGRAPH/Eurographics symposium on computer animation*, Vienna, Austria, Sep. 2006, pp. 251–260.

[6] 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.

[7] G. Hofer and K. Richmond, "Comparison of HMM and TMDN methods for lip synchronisation," in *Proc. Interspeech*, Makuhari, Japan, September 2010, pp. 454–457.

[8] M. Tamura, T. Masuko, T. Kobayashi, and K. Tokuday, "Visual speech synthesis based on parameter generation from HMM: Speech-driven and text-and-speech-driven approaches," in *Proc. ICASSP*, May 1998, pp. 3745–3748.

[9] 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.

[10] Naturalpoint, 2013. [Online]. Available: http://www.naturalpoint.com/optitrack/

[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, May 2012, pp. 3313–3316.

[12] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, Oct. 1996, pp. 1137–1140.

[13] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[14] 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.

[15] 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.