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.

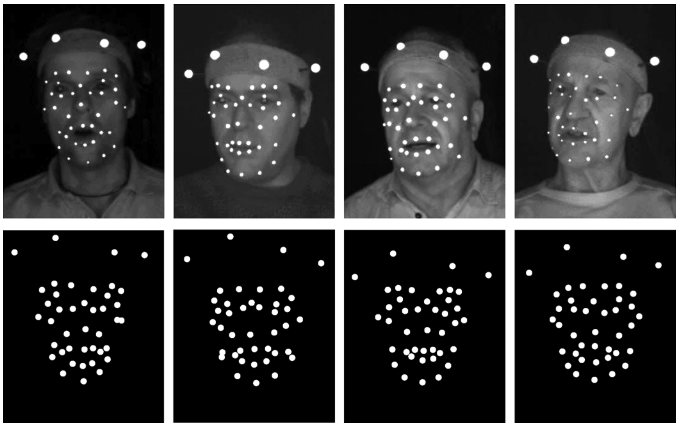

Fig. 1. Still images from grayscale videos showing facial marker layout (top) for four different speakers and corresponding renderings of 3D marker data (bottom).

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

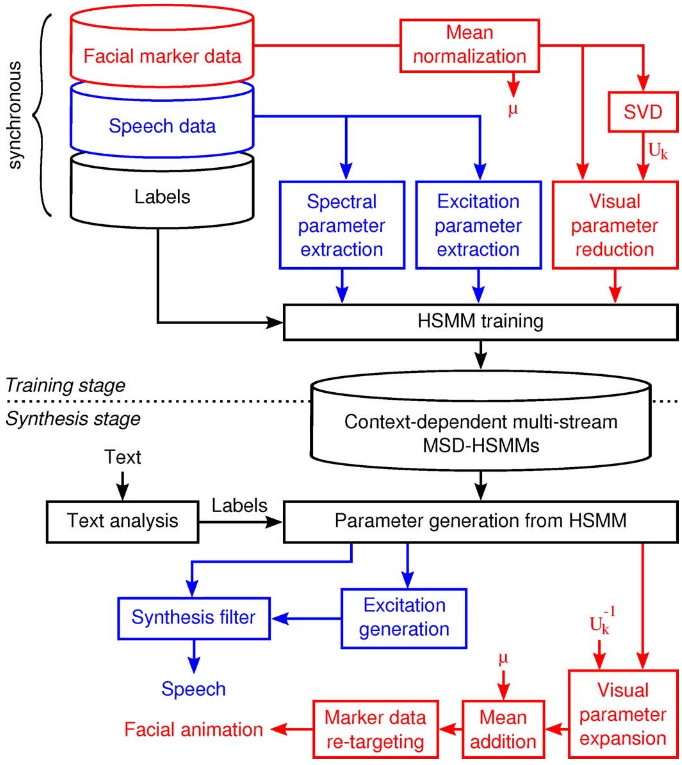

Fig. 2. Overview of a speaker dependent audiovisual speech synthesis system, which consists of three main components: audiovisual speech analysis, audiovisual training, and audiovisual speech generation. The corresponding audio-only system does not include the red parts, and the corresponding visual-only 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 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.

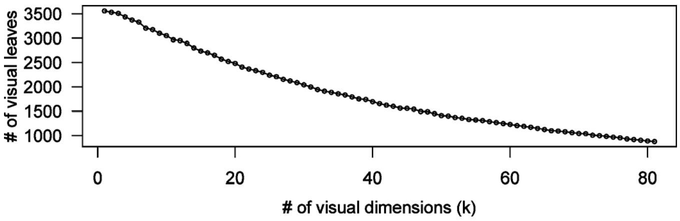

Fig. 3. Average number (across four speakers) of total leaf nodes in the visual clustering trees as a function of visual PCA dimensions kept (k).

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

$$d_A(k) = \mu_A(k) + \rho \cdot \sigma_A^2(k) \qquad \text{for } 1 \leq k \leq K, \tag{1}$$

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

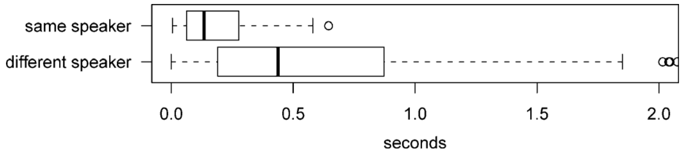

Fig. 4. Boxplots for the differences in utterance length between audio-only and visual-only synthesized utterances. For 23 test utterances and 4 speakers, the top boxplot contains all 92 combinations where audio and visual models were from the same speaker, and the bottom boxplot contains all 276 combinations 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

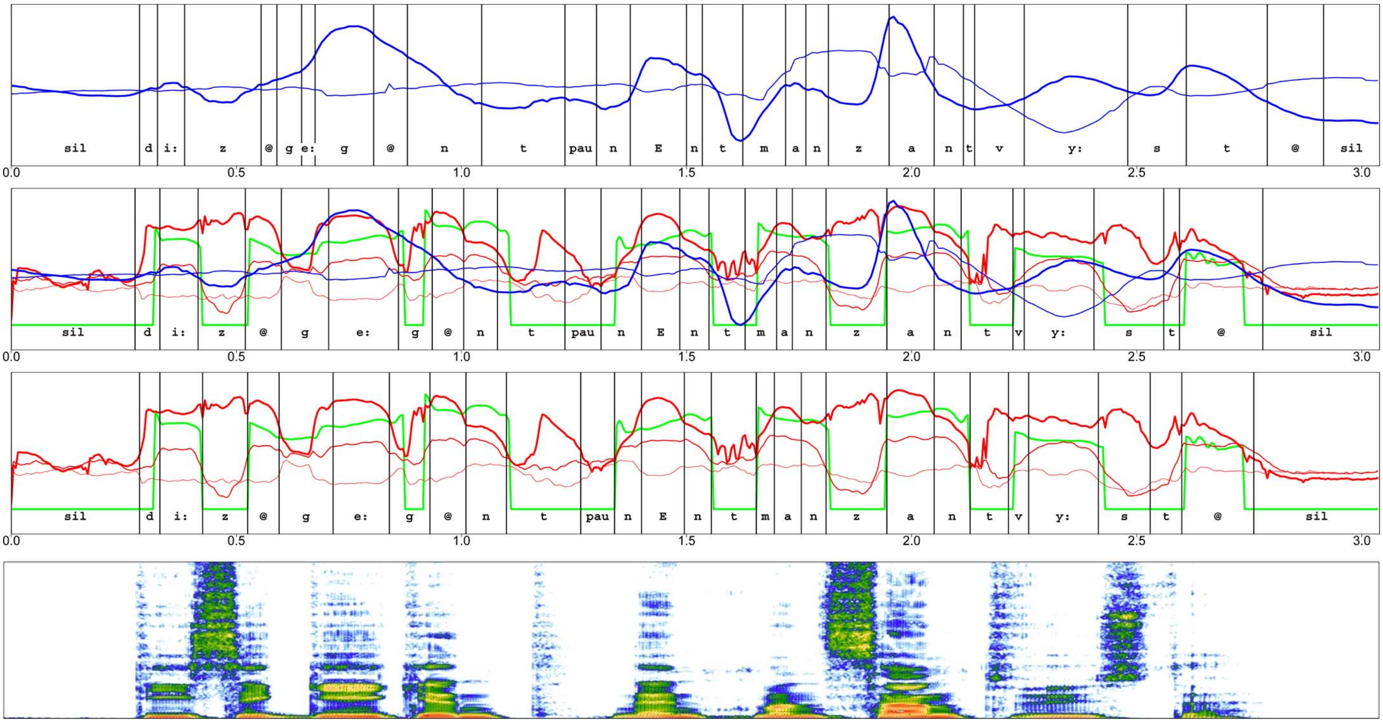

Fig. 5. Result of forced alignment using visual (first row), audiovisual (second row) and audio (third row) models and data. The plots show the Euclidean distance between the central upper lip and central lower lip markers (thick blue line), the Euclidean distance between the left and right mouth corner 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.

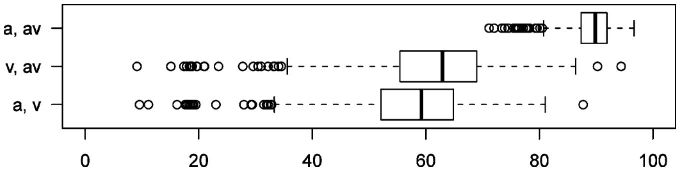

Fig. 6. Boxplots for matching percentage per utterance for audio and audiovisual models (top), visual and audiovisual models (middle) and audio and visual models (bottom).

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, ɡ, 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, ɡ, 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

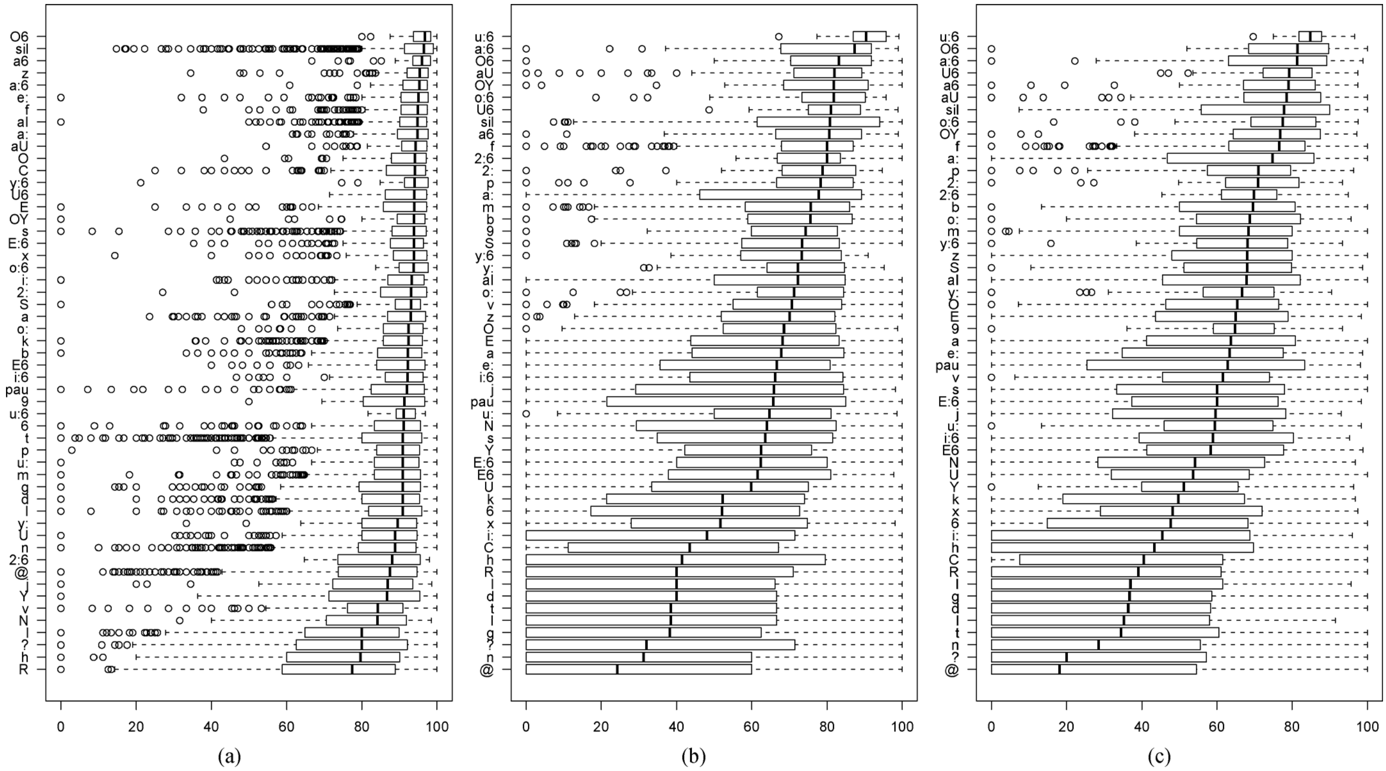

Fig. 7. Boxplots for matching percentage per phone. (a) Auditory and audiovisual. (b) Visual and audiovisual. (c) Auditory and visual.

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.

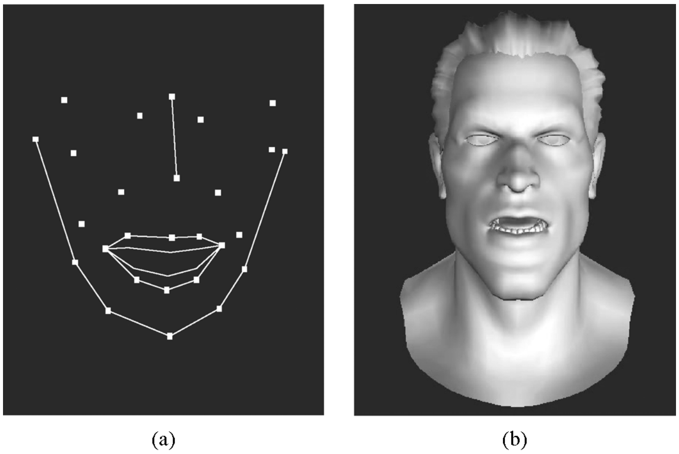

Fig. 8. Mode of speech motion presentation in the first and second (expert) evaluations. Example videos are available on http://userver.ftw.at/~schabus/jstsp2013. (a) Raw marker data. (b) Data-controlled 3D head.

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].

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	

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 "$\ast$" indicates statistical significance of the score difference according to Bonferroni-corrected Pearson's χ^2-tests of independence with $p < 0.05$.

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

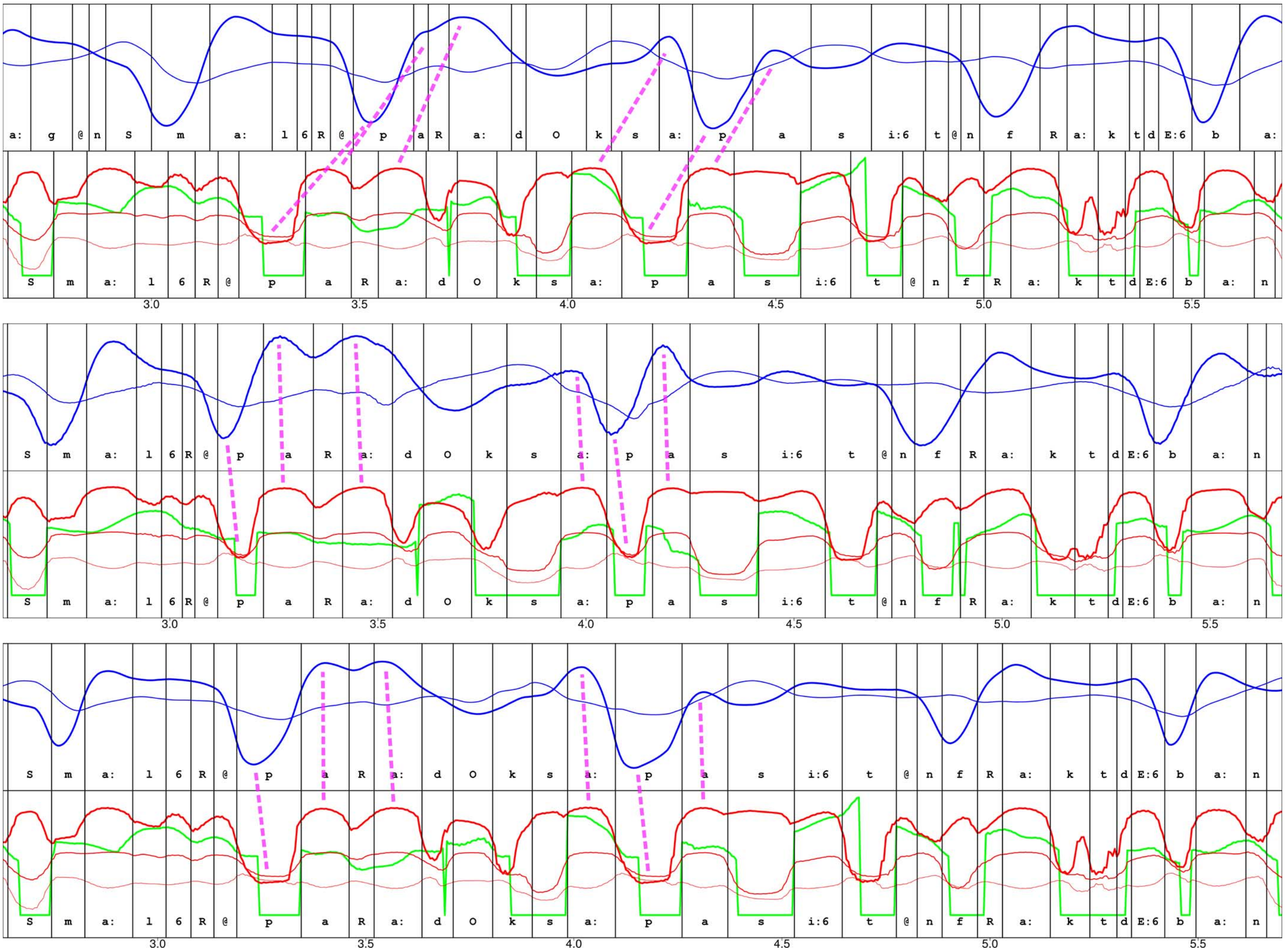

Fig. 9. Excerpts of synthesized audiovisual trajectories for one of the "challenging" utterances from different synthesis strategies: uttlen-visual (top), audiovisual (middle) and durcopy-audio (bottom). The plots show the Euclidean distance between the central upper lip and central lower lip markers (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

Compared Methods	experts			non-experts		
	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 illustrates 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 M to obtain a normalized $\bar{M}$, and compute the Singular Value Decomposition (SVD):

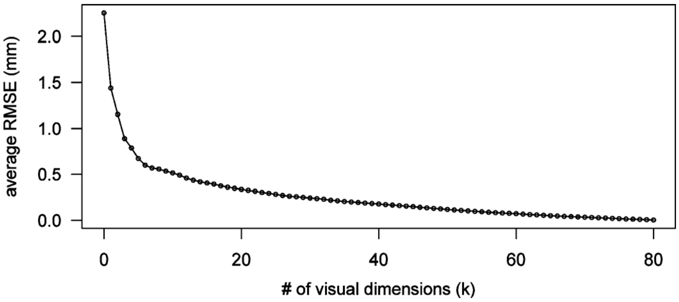

Fig. 10. Average (across four speakers) Root Mean Squared Error (RMSE) of four-fold cross validation PCA reconstruction of visual parameters with varying dimensionality (k).

$$\bar{M} = U \cdot \Sigma \cdot V^T \tag{6}$$

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.