VISUAL SPEECH SYNTHESIS FROM 3D MESH SEQUENCES DRIVEN BY COMBINED SPEECH FEATURES

Felix Kuhnke and Jörn Ostermann

Institut für Informationsverarbeitung, Leibniz Universität Hannover, Germany
{kuhnke,ostermann}@tnt.uni-hannover.de

ABSTRACT

Given a pre-registered 3D mesh sequence and accompanying phoneme-labeled audio, our system creates an animatable face model and a mapping procedure to produce realistic speech animations for arbitrary speech input. Mapping of speech features to model parameters is done using random forests for regression. We propose a new speech feature based on phonemic labels and acoustic features. The novel feature produces more expressive facial animation and it robustly handles temporal labeling errors. Furthermore, by employing a sliding window approach to feature extraction, the system is easy to train and allows for low-delay synthesis. We show that our novel combination of speech features improves visual speech synthesis. Our findings are confirmed by a subjective user study.

Index Terms— Visual Speech Synthesis, Facial Animation, Lip Synchronization, Speech Features

1. INTRODUCTION

During the last decades the synthesis of realistic talking virtual human faces has been a major concern of research. The goal is to produce visuals indistinguishable from real faces and further to produce linguistically correct speech animation. Recent advances in 3D facial performance capture allow automatic capture of very realistic facial geometry (mesh sequences), ideal for creating digital doubles. While visual speech synthesis has been based on various recordings, an efficient automatic solution based on captured 3D mesh sequences is still missing. Synthesis of visual speech is usually driven by a sequence of phoneme labels, but the results of automatic or manual phoneme labeling can be imprecise. Furthermore, phonemes only describe a fixed set of speech units, information about the individual acoustic presentation is lost.

To address these issues, this paper proposes an effective framework for visual speech synthesis. Our animation system can be driven by arbitrary speech features, but we suggest a novel combination to robustly handle inaccuracies in phoneme labeling and produce more expressive animations. Our system can be driven from text-to-speech output or real audio recordings using acoustic and phonemic descriptions of speech. We directly create our model from 3D performance capture data, and no manual modeling is required. We do not define or need a frame or motion dictionary for visual synthesis, such as visemes. Instead of using specifically hand tailored methods, we use an off-the-shelf regression method. Our contributions are as followed:

- We propose a new phonemic feature vector for facial animation and further show the benefits of combining different speech features.
- We show how 3D facial performance capture data can be used for visual speech synthesis with a regression based method and propose automatic phoneme-guided 3D mesh processing.
- We are the first to use a publicly available database [1] to synthesize 3D visual speech to make results comparable.

Furthermore, we show the effectiveness of our approach in a subjective user study. From the results it seems that our approach already works well from a small database of 40 sentences.

The paper begins by reviewing related work. Section 3 introduces the components of our framework. We explain how to use performance capture data and present the extraction of combined speech features. In Section 4 we describe the design and conduction of a subjective user study to evaluate our method and subsequently conclude our work.

2. RELATED WORK

Over the last years numerous visual speech synthesis systems have been proposed. A very detailed and recent review can be found in [2]. Highly related to visual speech synthesis is the field of audio-visual speech synthesis. These approaches (e.g. [3]) jointly synthesize auditory and visual speech from text input. In this work we focus on visual speech synthesis, to keep the synthesis of auditory speech and visual speech untangled.

The general task of visual speech synthesis is to provide a **mapping** from a given **auditory speech input**, possibly with

978-1-5090-6067-2/17/$31.00

additional information, to a visual speech animation using a **face model**.

Auditory speech input: Visual speech synthesis systems can be driven by categorical speech features such as phoneme labels or by continuous acoustic speech features such as Mel Frequency Cepstral Coefficients (MFCC) extracted from auditory speech input. To model coarticulation (see Section 3.4) phoneme driven systems often assume a temporal context and model speech by tri- or quinphones [4]. A simpler approach is to sample the phoneme labels at fixed time steps [5] for a given temporal context. Similarly, acoustic speech features such as MFCC and others (see [6] for a comparison) sampled at different time steps can be combined to yield one context dependent feature [4,6–8].Another widespread approach is to use MFCC delta features (e.g. [9]).

Phoneme labels assume a fixed dictionary of speech sounds. As a result, information about the individual acoustic presentation of a phoneme is lost. If a phonemic labeling is used to drive the animation, the labeling of the training and testing data is often produced or corrected by a human annotator (e.g. [10–12]) as automatic phoneme labeling can be inaccurate.

To our knowledge, the effects of using both, acoustic and phonemic speech features, have not been explored yet.

Face Models: Earlier approaches used 2D image-based rendering techniques to produce speech animations e.g. [13]. More flexibility is provided by 2D models, with the most used being active appearance models (AAMs) [14]. However, controllable 3D models have the benefit that they can be used to synthesize arbitrary head poses, lighting conditions and can be placed in any virtual environment. Recent advances in 3D facial performance capture demonstrate that high fidelity 3D capture of human facial appearance is possible [15,16]. However, there is no publicly available speech database from these recent capture systems. A 3D speech database was introduced by Fanelli et al. [1] but to date, no system has been proposed to directly synthesize visual speech based on their data.

Most 3D visual speech synthesis approaches use motion-capture data (a sparse set of points, tracked on the recording) to animate a predefined, manually created, face model (e.g. [17]), or animate a denser mesh by interpolating the dense vertices from a sparse set of captured vertices [8]. Another technique is 2D-to-3D reconstruction, where the source material is a 2D video wrapped to a 3D head [9].

Wampler et al. [12] and Müller et al. [18] generate face models from performance capture data or 3D scans of multiple persons. However, their main focus is on building multi-person models that can be adapted to (at least) a single input 3D mesh.

Mapping: Mapping speech features to model parameters can be done in various ways [2]. Typically Hidden Markov Models (HMMs) are used to predict the facial appearance from speech features.

Another approach, unit selection, selects appropriate samples from a database. Concatenation of original visual frames or subsequences like visemes, dynamic visemes [11], or animes [12, 17] produces a novel speech animation. The majority of systems assume such a fixed dictionary of visual speech units. Any concatenation of original recording data provides static realism. However, good synchronization and realistic motion of the concatenated sequence is not guaranteed. Besides, using any kind of visual dictionary requires us to generate it first (e.g. using clustering [11]).

Regression methods directly map speech features to visual appearance without assuming any fixed units of visual speech, neither during prediction nor during visual synthesis. We exclude HMMs here, as they internally work with states. Craig et al. [7] use multilinear regression to map multiple adjacent MFCCs to face model parameters. Neural Networks (NN) have been used, among others, by Theobald and Matthews [4] and Takacs et al. [19]. Recently, Kim et al. [5] proposed a general framework for spatiotemporal sequence prediction. They extract phoneme labels using a sliding window approach and use random forests [20] to estimate the parameters of an AAM to synthesize speech animations. Regression-based methods require less prior assumptions about how to model speech by assuming that a fixed-length temporal context is sufficient to model visual speech. Therefore we choose the regression approach as mapping procedure. Our work is related to the works of Kim et al. [5] but we extend their method to 3D performance capture data and a novel speech feature combination.

3. VISUAL SPEECH FROM FACIAL PERFORMANCE CAPTURE

3.1. Input data

We assume a given 3D mesh sequence which we define as a sequence of registered 3D face scans (meshes) of a person. The number of vertices N is constant and inter-frame vertex correspondence is known for all frames M. While obtaining such data is a challenging task on its own, we leave it to the performance capture community. Furthermore, we assume a corresponding audio recording for every 3D mesh sequence. Every audio recording has a phonemic labeling. Phoneme labels can be produced automatically from a speech transcript or automatic speech recognition using forced alignment techniques.

3.2. Face model

Using the aligned meshes directly as visual model would be computational unfeasible. We follow the common approach to create a decomposition of the geometric vertex points into a component model. To compute a linear component face model we perform Principle Component Analysis (PCA) on the aligned meshes. We can, however, without any change of method, switch to a different parameter-driven model, such

as a blendshape model. The PCA model allows us to remove components with low explanatory value, which will greatly reduce the computational burden required in the following steps. Using PCA can even remove noise in the performance capture data, as we encountered in Section 4. Projecting our training meshes $V_{i=1,\ldots,M} \in \mathbb{R}^{N3}$ into the truncated PCA space, we obtain low dimensional visual parameter vectors $Y_{i=1,\ldots,M} \in \mathbb{R}^{D}$ for all meshes of our recording.

3.3. Visual data preparation

Visual speech should be captured with high frame rates to capture the subtle motions of speech and the fine temporal dependence between auditory and visual speech. If high frame recordings are not available, upsampling can improve the synthesis results. As we now have a sequence of Ys we interpolate between them using cubic splines, to produce new intermediate frames. To ensure bilabial mouth closure we correct the interpolation using a bilabial constraint (see below). These frames might not replace original high frame rate recordings, but for us, improved synthesis quality non the less.

Unvoiced parts of the recordings provide no speech features and wide variations of facial movements such as breathing and facial gestures. These parts can be simply replaced by a neutral face of the subject.

Bilabial constraint: Bilabial (p,b,m) mouth closure can be lost due to model smoothing during performance capture and/or low recording frame rate. Phonemic labels can be used to restore mouth closure for bilabials. To do so, we need to measure the mouth closure of our mesh. One way is to locate two vertices that represent upper and lower lip. This can be done either manually or using a facial feature point detector. The Euclidean distance between a centered lower lip vertex and a centered upper lip vertex is our closure measure. Now, we need to search for local closure maxima near frames with bilabials as phoneme labels. To force mouth closure, we calculate the parameter directions of closure G from our visual parameters around the frame with maximum closure found at index t_m with

$$G = Y_{t_m} - \frac{1}{2}(Y_{t_m-1} + Y_{t_m+1}). \quad (1)$$

Then we search for the weighted amount of G that needs to be added to Y_{tm} to achieve full mouth closure. Smoothing this change over the neighboring frames provides more realistic results. This idea can be extended, e.g. to enforce protrusion for certain phonemes.

3.4. Speech feature extraction

The acoustics of speech are classically modeled with phonemes. Phonemes describe a fixed dictionary of sounds to produce speech. In phonology the concept of **allophones** describes that a single phoneme can be pronounced in different ways, resulting in different visual and acoustical versions of the same phoneme. This effect is tightly linked to **coarticulation**, the effect that the preceding and following speech influence the current articulation. The effect of coarticulation is dependent on the sequence of phonemes and speech rate.

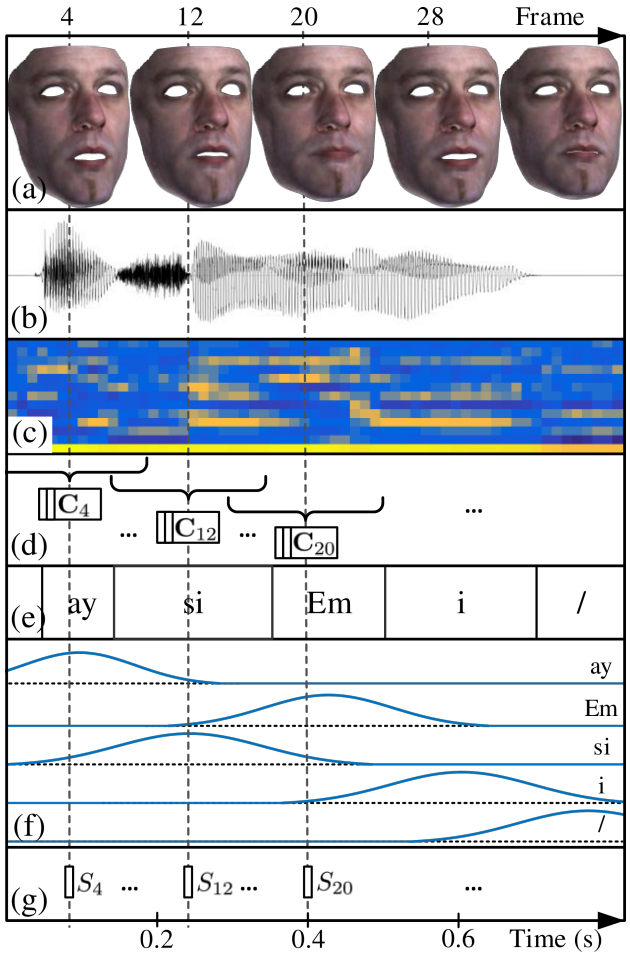

Fig. 1. Input data and feature extraction: The data used in our speech synthesis system is shown in the lanes a to g with exemplary feature extraction for frames 4,12 and 20. Lanes: (a) Textured mesh sequence. (b) Audio waveform. (c) MFCCs. (d) MFCC feature matrices C extracted using a sliding window. (e) Phoneme labels. (f) Smooth phoneme signals. (g) Smooth phoneme feature vectors sampled from (f).

Acknowledging both phenomena we conclude that visual speech needs to be derived from an auditory context, where the local characteristics loudness, duration and spectral shape of speech are considered. These characteristics are contained in acoustic speech features such as MFCCs. Similar to previous works we therefore extract acoustic features from a sliding window of length L around the corresponding visual frame. We use MFCCs on overlapping subsequences of the

analysis window. The resulting MFCCs are concatenated to yield one context feature matrix **C** per frame, similar to a spectrogram. The process is visualized in Figure 1, lanes (c) and (d).

Smoothed phoneme feature vector: Despite the mentioned limitations, phoneme labels have the advantage that they are usually generated using language models, i.e. the labeling process incorporates knowledge of the language. This makes them much more robust than acoustic features to noise or mumbling and ambiguity errors. We therefore expect an improvement if phonemic features are added to the acoustic feature matrix.

We could sample the phonemic labels with the same sliding window procedure as for the MFCC features, to obtain a representation such as in [5]. The resulting vector P has categorical entries, and the temporal resolution is fixed to the sampling frequency. We therefore propose a novel phonemic feature that models the phoneme context as a continuous vector, yielding a smoothed representation of the current phoneme context. We explicitly blur the fixed temporal information that is encoded in sequential phoneme features and aim for a representation that describes current phoneme probabilities.

To obtain our feature vector we assume that every phoneme has a temporal center, lying at the center between the beginning and end time of the phoneme. Every phoneme has a symmetrical context window around this center, with the length of the phoneme plus an additional fixed length J. The additional length J enables to model the coarticulation and the temporal uncertainty of phonemic labeling (see [21] for research on labeling errors). To encode temporal information, we use a Gaussian window with standard deviation $\sigma = 0.4$. The smoothed phoneme vector (SPV) can then be generated for any point in time by sampling the values of the windows for every phoneme. In effect, the SPV, named S, has the dimension of the size of phonemes in the dictionary. If multiple windows of the same phoneme overlap, the maximum value is used to keep the information for the phoneme with highest influence. The smoothed phoneme signal and exemplary sampling of SPVs is illustrated in Figure 1, lanes (f) and (g). To summarize, the SPV provides a snapshot of the local phonemic context and provides temporal information in the magnitudes.

Further benefits of the continuous representation are that techniques, such as neural networks do not work with categorical input and require an encoding of categorical variables. As it is desirable to keep the feature dimension low, SPV only has the dimension of the phoneme dictionary, whereas a full one-hot encoding would be dictionary size times the number of samples per context.

Concatenating and flattening features C and S produces a speech feature vector X for every frame in the database.

3.5. Regression

The problem of facial animation is now reduced to a regression problem, namely finding Y for a given X, and more precisely to find a function $h(X) := Y$. In practice, the goal is to find a predictor h that minimizes some loss $l(h(X), Y)$ over a training set. In our case we wish to learn a predictor that maps an input speech feature X to a visual parameter vector Y. Because Y is multidimensional, we use the squared Frobenius norm and define our loss function as

$$l(h(X), Y) = \|h(X) - Y\|^2_{Fro} \tag{2}$$

The regression task is to find an h which minimizes the loss over our training database $\{(X_i, Y_i)\}_{i=1}^{M}$.

At this point we can use generic off-the-shelf regression techniques, including general linear models, neural networks and random forests. As proposed in [5] we choose random forests to solve the regression task. Using random forests has the added benefits that it can handle categorical and continuous covariates, supports multi-output regression and that training of the trees can be done in parallel.

3.6. Synthesis and post-processing

Using the trained predictor, we are now able to synthesize the face model parameters for any given sequence of speech features. The results are already good, but further post-processing can improve the perceived quality. We use our bilabial constraint technique again in post-processing, to ensure full mouth closure in the animation at the appropriate phonemes. In visual speech synthesis filtering or blending is usually performed after the synthesis step, to counteract jittery animation (e.g. [17] [12]). We stick to a parameter-wise filtering as proposed by Cao et al. [17]. A low-pass filter is applied parameter-wise with cut-off frequencies learned from the training data.

As a final step, the face model parameters are projected back to vertex space for rendering.

4. EVALUATION

Methods: The most widely used methods to measure the objective quality of visual speech is to compare synthesized animation parameters or the geometric model to the recorded ground truth data. However, it is still an open issue how the objective measures can be used to to give a reliable indication of subjectively perceived quality [4]. As subjective evaluation is still the most significant measure, we conduct a subjective user study to evaluate the quality of our animations.

Data: To test our method we use the Biwi 3D Audiovisual Corpus of Affective Communication [1]. It includes speech of 14 different subjects, citing 40 sentences once emotional and once neutral. Registered 3D mesh sequences at 25 fps and phonemic labels of the audio sequences are provided with

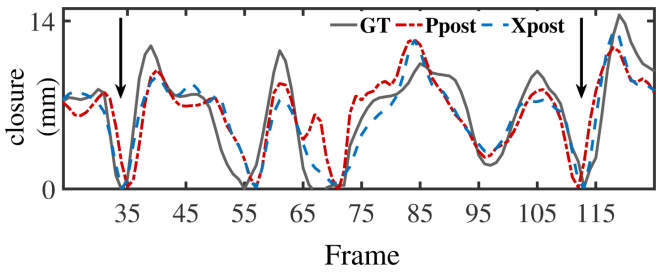

Fig. 2. Comparison of synthesized mouth closure for Xpost (proposed), Ppost and GT. Notice the improved temporal alignment to GT obtained with Xpost at frames 34 and 113.

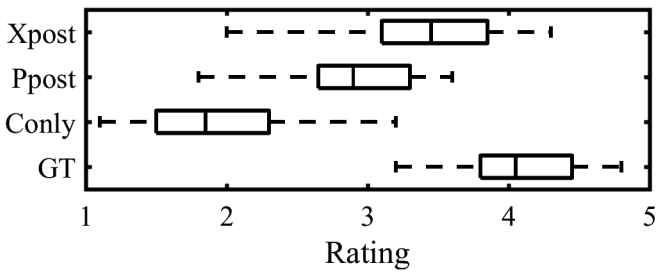

Fig. 3. Boxplot of the subjective results.

the database. As we are not interested in emotional speech nor multi-person models, we only use the neutral sentences of male one for our test.

Parameter tuning: Even though we strive for a full automatic system, one needs to set certain parameters, such as the size of the sliding windows. While the best settings are generally unknown, we use optimal parameters with respect to the loss function (2) as objective quality measure. We found a sliding window L=320 ms to extract the acoustic feature C, similar to the 334 ms context window found by [6]. SPV features S are generated with a window J=280 ms. Our windows are symmetric around the center frame.

MFCCs are calculated with a window size of 40 ms in steps of 20 ms. Our PCA space has D=17. We upsampled the 25 fps sequences to 50 fps using the bilabial constraint technique.

User study: In our subjective user study we let 20 human subjects rate the synthesized animations of 10 test sentences. The sentences are generated by training 10 predictors in a leave-one-out fashion for every method under comparison. Furthermore, we add the ground truth sequence. We present the animations in random order for every sentence and subjects could repeat animations. For the rating we stick to the common realism scale approach [3, 19]. Subjects had to rate the perceived naturalness of the speech animation on a 1-5 scale. Where 5 corresponds to "very realistic", natural speech motion and 1 to "completely unreal". As we do not provide any appearance for inner mouth or eyes, the subjects were instructed to rate the face and lip movements, instead of complete facial appearance.

We perform our evaluation for different setups of our synthesis system switching between different features and post-processing. Approach "GT" just uses the ground truth recordings. In a preliminary study subjects where disturbed by jittery motions in the original data. For the final study we projected the ground truth meshes to our truncated PCA space to smooth jittery capture results. "Xpost" uses our combined feature X with post-processing. "Conly" is driven by Cs only, without post-processing. We omit post-processing, to include one setup that does not use any phoneme based features nor processing steps during synthesis. "Ppost" uses categorical Ps obtained as in [5].

Results: The boxplot in Figure 3 shows that our approach Xpost was rated with second highest median realism (3.45) after the ground truth sequences (4.05). The Conly approach was rated with low realism (1.85), which we think is partly because it did not benefit from the forced closure post-processing. Furthermore, Ppost (2.9) did not perform as well as Xpost (3.45). In our experience, this is caused by the imprecise phoneme labeling of the test sequences. With Xpost, the regression can utilize the precise temporal information from the MFCC features and additional phoneme context from the SPV features. The improved temporal alignment can be seen in Figure 2. Furthermore, the animation is more expressive, as acoustic information such as pitch and loudness through MFCCs are available to the predictor. However, the Ppost and Xpost boxes overlap and as indicated by a wide range of ratings the level of agreement between subjects seems to be low. Nonetheless, 75% of subjects rated Xpost higher than the Ppost, suggesting a higher realism with Xpost compared to Ppost.

An example video clip showing synthesized results used in the evaluation is provided in the supplemental material[1]. The example clip shows the weaknesses of Ppost and Conly approaches compared to Xpost.

We admit that in our case realism scores can only be seen as approximative, as we did not animate a full face. On the other hand, eyes and teeth are rigid objects that can be added to the face in an additional step. A tongue model could be animated with a similar system as proposed here.

5. CONCLUSION

We have demonstrated a framework to build a visual speech synthesis system from 3D performance capture data using a publicly available 3D database. We suggested refinement methods to pre- and post-process visual data including phonemic constraints. We introduced a novel phonemic feature, a smoothed continuous description of the phonemic context. Furthermore, we proposed to combine phonemic and acous-

[1] http://www.tnt.uni-hannover.de/projects/facialanimation/icme2017

tic speech features to drive facial speech animation. A user study confirmed that our novel combination outperformed traditional features using a regression-based system to create facial speech animation. The proposed method is a step towards fulfilling our vision to automatically create versatile talking avatars from a small set of recordings of a person.

6. REFERENCES

[1] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-D Audio-Visual Corpus of Affective Communication. *IEEE Transactions on Multimedia*, 12(6):591–598, October 2010.

[2] Wesley Mattheyses and Werner Verhelst. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66:182–217, February 2015.

[3] Robert Anderson, Bjorn Stenger, Vincent Wan, and Roberto Cipolla. Expressive Visual Text-To-Speech Using Active Appearance Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3382–3389. IEEE, June 2013.

[4] Barry-John Theobald and Iain Matthews. Relating Objective and Subjective Performance Measures for AAM-Based Visual Speech Synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2378–2387, October 2012.

[5] Taehwan Kim, Yisong Yue, Sarah Taylor, and Iain Matthews. A Decision Tree Framework for Spatiotemporal Sequence Prediction. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 577–586. ACM, ACM Press, 2015.

[6] Praveen Kakumanu, Anna Esposito, Oscar N. Garcia, and Ricardo Gutierrez-Osuna. A comparison of acoustic coding models for speech-driven facial animation. *Speech Communication*, 48(6):598–615, June 2006.

[7] Matthew S. Craig, Pascal van Lieshout, and Willy Wong. A linear model of acoustic-to-facial mapping: Model parameters, data set size, and generalization across speakers. *The Journal of the Acoustical Society of America*, 124(5):3183, 2008.

[8] Lucas Terissi, Mauricio Cerda, Juan C. Gomez, Nancy Hitschfeld-Kahler, Bernard Girau, and Renato Valenzuela. Animation of generic 3d head models driven by speech. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pages 1–6. IEEE, 2011.

[9] Lijuan Wang, Wei Han, Frank K. Soong, and Qiang Huo. Text Driven 3d Photo-Realistic Talking Head. In *INTERSPEECH*, pages 3307–3308, 2011.

[10] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. JALI: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics*, 35(4):1–11, July 2016.

[11] Sarah L. Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. Dynamic units of visual speech. In *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pages 275–284. Eurographics Association, 2012.

[12] Kevin Wampler, Daichi Sasaki, Li Zhang, and Zoran Popović. Dynamic, expressive speech animation from a single mesh. In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 53–62. Eurographics Association, 2007.

[13] Kang Liu and Joern Ostermann. Realistic facial expression synthesis for an image-based talking head. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, July 2011.

[14] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 23(6):681–685, 2001.

[15] Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics*, 34(4):46:1–46:9, July 2015.

[16] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of Personalized 3d Face Rigs from Monocular Video. *ACM Transactions on Graphics (TOG)*, 35(3):28, 2016.

[17] Yong Cao, Wen C. Tien, Petros Faloutsos, and Frédéric Pighin. Expressive speech-driven facial animation. *ACM Transactions on Graphics (TOG)*, 24(4):1283–1302, 2005.

[18] P. Müller, G.A. Kalberer, M. Proesmans, and L. Van Gool. Realistic speech animation based on observed 3-D face dynamics. *IEE Proceedings - Vision, Image, and Signal Processing*, 152(4):491–500, 2005.

[19] György Takács. Direct, modular and hybrid audio to visual speech conversion methods-a comparative study. In *INTERSPEECH*, pages 2267–2270, 2009.

[20] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[21] John Kominek, Christina L. Bennett, and Alan W. Black. Evaluating and correcting phoneme segmentation for unit selection synthesis. In *INTERSPEECH*, 2003.