PHOTOREALISTIC ADAPTATION AND INTERPOLATION OF FACIAL EXPRESSIONS USING HMMS AND AAMS FOR AUDIO-VISUAL SPEECH SYNTHESIS

Panagiotis P. Filntisis, Athanasios Katsamanis and Petros Maragos

School of ECE, National Technical University of Athens, 15773 Athens, Greece
Athena Research and Innovation Center, 15125 Maroussi, Greece
filby@central.ntua.gr, {nkatsam, maragos}@cs.ntua.gr

ABSTRACT

In this paper, motivated by the continuously increasing presence of intelligent agents in everyday life, we address the problem of expressive photorealistic audio-visual speech synthesis, with a strong focus on the visual modality. Emotion constitutes one of the main driving factors of social life and it is expressed mainly through facial expressions. Synthesis of a talking head capable of expressive audio-visual speech is challenging due to the data overhead that arises when considering the vast number of emotions we would like the talking head to express. In order to tackle this challenge, we propose the usage of two methods, namely Hidden Markov Model (HMM) adaptation and interpolation, with HMMs modeling visual parameters via an Active Appearance Model (AAM) of the face. We show that through HMM adaptation we can successfully adapt a "neutral" talking head to a target emotion with a small amount of adaptation data, as well as that through HMM interpolation we can robustly achieve different levels of intensity for an emotion.

Index Terms— expressive audio-visual speech synthesis, photorealistic, hidden Markov model, interpolation, adaptation

1. INTRODUCTION

Synthesis of talking heads constitutes a field that is enjoying an increasing amount of focus due to their usage in plentiful applications that pertain to Human-Computer Interaction (HCI). A vital part of making these interactions feel as natural as possible for the human interlocutor is the ability of the talking head to express emotions [1]. Facial expressions are tightly coupled with emotions [2], and expressive behavior itself is a vital part of social life. In [3] it was found that the visual channel of information (i.e., the face and its expressions), played a more important role than the acoustic and verbal information, in regards with the perception of social signals.

If we consider the vast number of facial expressions/emotions the talking head must be able to express in order to achieve successful social interaction with humans, it is obvious that the amount of required training data increases exponentially. To tackle this, we desire two abilities for the talking head to possess: (a) the ability to express a new emotion given a small amount of corresponding data and (b) the ability to express a specific emotion in continuous intensity levels and form more complex emotions through combinations of emotions. We can see the different intensity levels of an emotion as its combination with a "neutral" emotion. Our motivation for the aforementioned abilities stems from a number of studies on emotion that have shown both that emotions have different intensity levels [4, 5] and that emotions can be combined in order to form more complex expressions [5].

In this paper, in the context of parametric photorealistic audio-visual text-to-speech synthesis (AVTTS) using Hidden Markov Models (HMM), where HMMs model both acoustic and visual parameters (through an Active Appearance Model - AAM), we study and propose the usage of two different methods for providing the resulting talking head with the two aforementioned abilities. In order to adapt a "neutral" talking head to a target emotion with a low amount of adaptation data we employ HMM adaptation [6]. In order to synthesize expressive speech of different intensity levels and combinations of expressions we use HMM interpolation [7].

In the field of parametric expressive TTS synthesis, HMM interpolation and adaptation have been successfully used in [8], for obtaining acoustic synthetic speech with intermediate emotional styles, and for adapting a neutral HMM-based acoustic text-to-speech system to the emotions of joy and sadness. Previous approaches on parametric expressive AVTTS include both graphics based models and photorealistic models. In graphics based approaches, facial expressions are modeled from a set of images using a set of parameters such as MPEG-4 Facial Animation Parameters, which are then used to drive a 3D model [9, 10]. Photorealistic expressive AVTTS was explored in [11, 12] where expressions and mixtures of expressions were modeled via cluster adaptive training.

The rest of the paper is organized as follows: We first describe the features used for the generation of the talking head though AAMs. We then give an overview of HMM-based AVTTS and describe the HMM adaptation and interpolation methods. Next we present our experimental results and the last section includes our concluding remarks.

2. ACTIVE APPEARANCE MODEL FEATURES FOR EXPRESSIVE AUDIO-VISUAL SPEECH SYNTHESIS

We model the face by employing an independent Active Appearance Model [13, 14]. In independent AAMs the shape and texture of the face are modeled separately. The shape s is defined by the coordinates of K vertices that outline the object being modeled. By obtaining the mean shape $\bar{\boldsymbol{s}}$ over M manually labeled images (the "training set") of the object, that have undergone a Procrustes analysis [15], we can express the shape of a particular image as the following linear combination:

$$\boldsymbol{s} = \bar{\boldsymbol{s}} + \sum_{i=1}^{n} p_i \boldsymbol{s}_i \tag{1}$$

where $\boldsymbol{s}_i$ are the n eigenvectors (called eigenshapes) calculated after employing a Principal Component Analysis (PCA) to the

This work has been funded by the BabyRobot project, supported by the EU Horizon 2020 Programme under grant 687831.

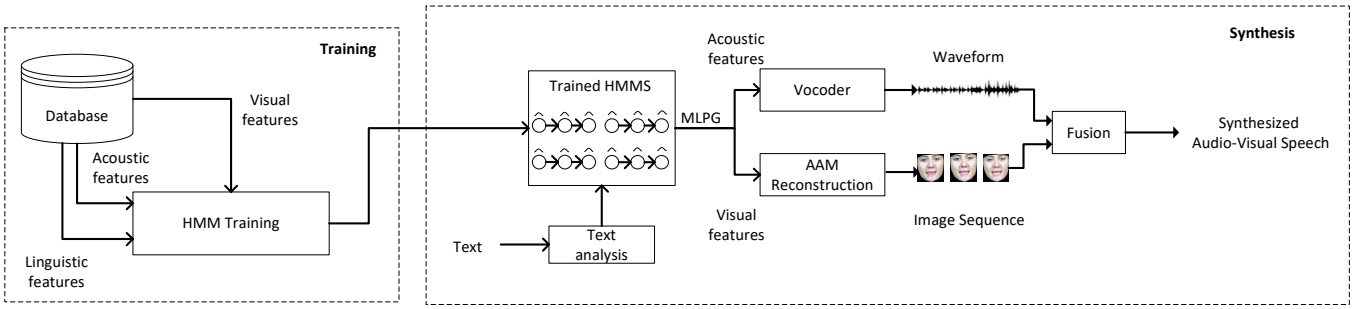


Fig. 1: HMM-based audio-visual speech synthesis.

shapes of training set and p_i is the weight that multiplies the i-th eigenshape.

In a similar way, if we define as $\boldsymbol{A}(\boldsymbol{x})$ the texture of the face, which is defined at the pixels $\boldsymbol{x}$ that lie inside the mesh of the vertices of the mean shape $\bar{\boldsymbol{s}}$, we can express the texture of a particular facial image as the linear combination:

$$\boldsymbol{A}(\boldsymbol{x}) = \bar{\boldsymbol{A}}(\boldsymbol{x}) + \sum_{i=1}^{l} \lambda_i \boldsymbol{A}_i(\boldsymbol{x}) \tag{2}$$

where $\bar{\boldsymbol{A}}(\boldsymbol{x})$ is the mean texture that is found after *shape normalizing* each image by warping it according with the warp defined by the shape $\boldsymbol{s}$ and the mean shape $\bar{\boldsymbol{s}}$, $\boldsymbol{A}_i(\boldsymbol{x})$ are the l eigenvectors (called eigentextures) obtained after applying PCA to the *shape normalized* images, and λ_i is the weight multiplying the i-th eigentexture. An instance of the face can be reconstructed if we warp the texture $\boldsymbol{A}(\boldsymbol{x})$ from the mean shape $\bar{\boldsymbol{s}}$, to the calculated shape $\boldsymbol{s}$.

In order to find the parameters of an AAM that best express a novel image of the face, we seek to minimize the *error image* $E(\boldsymbol{q})$ where $\boldsymbol{q}$ is the concatenated shape and texture weights vector. The error image is defined as the difference between the image texture warped on the mean shape $\bar{\boldsymbol{s}}$ and the synthesized texture $\boldsymbol{A}(\boldsymbol{x})$. We also include a quadratic penalty that corresponds to a Gaussian prior [16] which regularizes the solution and the robustness of the fitting, especially in databases where the samples vary greatly. The error is then minimized using an iterative method. In our case we use the Variable Order Template Update Inverse Compositional algorithm described in [16]. In Figure 2 we show an example of the first eigentexture of an AAM, trained on an expressive corpus.

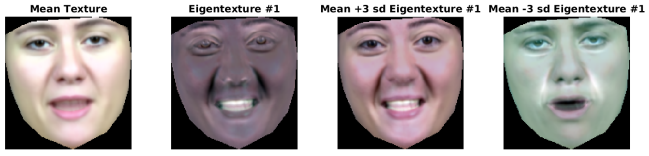


Fig. 2: Example of the first eigentexture and the variations it causes to the mean texture between values $[-3\sqrt{\lambda_1}, +3\sqrt{\lambda_1}]$ where λ_1 is the corresponding weight, for an AAM trained on an expressive corpus.

3. HMM-BASED AUDIO-VISUAL SPEECH SYNTHESIS

3.1. Overview

To synthesize audio-visual speech, we enhance the HMM-based acoustic speech synthesis architecture of [17] for audio-visual speech synthesis as depicted in Figure 1. Eigenshape and eigentexture weights are concatenated with acoustic features on an acoustic frame level, and the complete vector is used to train Multi-Space Distribution Hidden Semi-Markov Models (MSD-HSMMs) [18, 19]. First and second order differences of the features are included, so as to avoid discontinuities in the generated speech. This joint modeling of acoustic and visual features with multiple streams enforces a strong temporal alignment of the generated acoustic and visual streams. Training of the HMMs is done via the Baum-Welch algorithm and a decision tree based context clustering [20], based on the Minimum Description Length criterion [21], is used to obtain clusters of similar phonetic contexts for each different HMM stream.

At the synthesis stage, the text to be synthesized is analyzed in order to obtain its linguistic representation, and then a sentence HMM is constructed by the individual trained HMMs. Subsequently, smooth trajectories of acoustic and visual features are generated using the Maximum Likelihood Parameter Generation algorithm [22]. The waveform is generated from the acoustic features via the vocoder, while the visual features are used to reconstruct the image sequence through the trained AAM model of Section 2.

3.2. HMM Adaptation

In order to adapt an HMM-based AVTTS system to a target emotion we use the CSMAPLR (Constrained Structural Maximum a Posteriori Linear Regression) approach of [6].

Using the CSMAPLR adaptation, the state-output probability distributions for the target emotion are obtained by linearly transforming the mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of the trained models using a transformation matrix $\boldsymbol{Z}$ and a transformation bias $\boldsymbol{\epsilon}$:

$$\bar{\boldsymbol{\mu}} = \boldsymbol{Z}\,\boldsymbol{\mu} + \boldsymbol{\epsilon}, \qquad \bar{\boldsymbol{\Sigma}} = \boldsymbol{Z}\,\boldsymbol{\Sigma}\,\boldsymbol{Z}^T \tag{3}$$

The transforms are calculated using a recursive Maximum a Posteriori (MAP) criterion [23], which results in robust estimations even with a low amount of adaptation data. After calculating a global transform using all of the adaptation data, a regression tree structure that takes into account linguistic information is created and information about the prior distribution used in the MAP criterion is propagated from the parent nodes towards the leafs.

3.3. HMM Interpolation

By HMM interpolation, we mean using a combination of two or more HMMs in order to obtain an HMM with intermediate voice characteristics, assuming that all HMMs have the same topology. In [7], three methods are proposed for HMM interpolation: (a) interpolation between observations, (b) interpolation between output distributions of HMM states and (c) interpolation based on the Kullback-Leibler divergence.

In this paper, as in [24] we use the simplest of the three methods, interpolation between observations. More specifically, if we denote as $\boldsymbol{\mu}$ the mean and $\boldsymbol{\Sigma}$ the covariance matrix of the interpolated Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of an HMM state, we have:

$$\boldsymbol{\mu} = \sum_{i=1}^{K} \alpha_i \boldsymbol{\mu}_i, \qquad \boldsymbol{\Sigma} = \sum_{i=1}^{K} \alpha_i^2 \boldsymbol{\Sigma} \tag{4}$$

where K is the number of HMM models considered for the interpolation, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ is the mean and covariance matrix of the Gaussian distribution of the i-th model, and a_i is its corresponding weight. The weights satisfy the property: $\sum_{i=1}^{K} |\alpha_i| = 1$.

A problem that arises when considering HMM interpolation between models of HMM systems trained on different datasets is their different tying structures. In order to tackle this problem, we apply this interpolation method on the synthesis stage, by interpolating the Gaussian distributions of the states of the K different constructed sentence HMMs corresponding to the K different HMM systems.

4. EXPERIMENTAL RESULTS

4.1. Corpus

To evaluate the methods that we describe in this paper, we collected a corpus of audio-visual speech in four different emotions: neutral, happiness, anger, and sadness. A professional actress was hired to say 900 sentences in the Greek language for each of the aforementioned emotions. The actress was instructed to express each emotion in the most extreme manner. The captured footage includes video at 30 frames per second in $1080p$ resolution and audio from a high quality microphone at 44100 Hz sampling rate. In order to split the captured footage in sentences and obtain the phonetic alignment of the sentences we used the sail-align toolkit [25]. Table 1 shows the statistics on the post-processed corpus, which we call CVSP-Expressive Audio-Visual Corpus (CVSP-EAV), where some sentences had to be discarded due to equipment malfunctions.

Table 1: Statistics of the post-processed CVSP-EAV corpus.

Emotion	Neutral	Anger	Happiness	Sadness
Sentences	899	898	896	894
Duration	72 *min.*	71 *min.*	72 *min.*	86 *min.*
Frames	$\sim$129,000	$\sim$129,000	129,000	150,000

The acoustic features were extracted using the STRAIGHT toolkit [26] and include 31 mel-frequency cepstral coefficients, 25 band-aperiodicity coefficients, and the fundamental frequency, after resampling the audio at 16kHz (reconstruction from the features is done via the STRAIGHT vocoder). In order to extract the eigentexture and eigenshape weights associated with each frame we trained an AAM using 981 frames, spread among all four emotions, manually labeled at 61 facial landmarks. In the trained model, we kept the eigenshapes and eigentextures that account for 95% of shape and texture variations.

We then proceeded to fit the AAM to all frames of the database. In order to avoid artifacts that might appear if the fitting is not successful at a frame, we discarded sentences where the reconstruction error $E(\boldsymbol{q})$ was above 0.0030 for more than 10 frames of the sentence, as well as sentences where the reconstruction error averaged over all sentence frames was above 0.0018. We found heuristically the value 0.0018 to represent excellent fitting. The final size number of sentences that were used for training was 764 for each emotion.

4.2. Evaluation of HMM adaptation

After collecting the acoustic and visual features from the corpus, we first trained 4 different HMM-based AVTTS systems using the four different emotions of the CVSP-EAV corpus, which we will call the emotion-independent HMM systems. For training the HMM systems we used the HTS toolkit [17]. Next, we adapted the neutral HMM system to the other three emotions using a variable number of adaptation sentences: 5, 10, 20, 50, and 100.

Subjective evaluation of HMM adaptation was done via a web-based questionnaire. For each of the adapted HMM sets we generated 6 unseen sentences and each human evaluator was presented with 2 random videos from the 6 and was asked to evaluate the achieved expressiveness of the talking head on a scale of 1 to 3. In order to set a ground truth and avoid interpersonal differences, we also included for each adapted sentence, the same sentence generated from the corresponding HMM emotion-independent system as a ground truth value of 3. The results from the 32 respondents of the questionnaire can be seen in Figure 3.

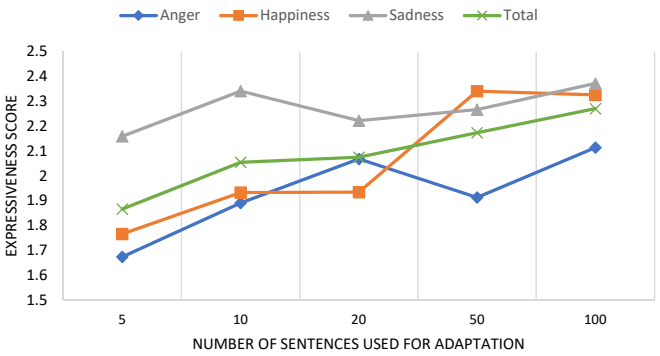
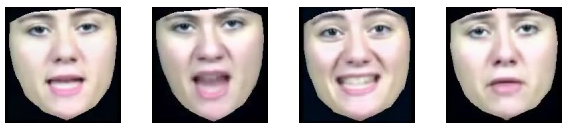


Fig. 3: Subjective evaluation results for the level of expressiveness achieved by a talking head, for a variable number of sentences, and for the three emotions of anger, sadness, happiness, as well as over all emotions.



(a) (b) (c) (d)

Fig. 4: Example frames from the same sentence: (a) a neutral HMM-based AVTTS system, and the same system adapted to (b) anger, (c) happiness, and (d) sadness, using 50 adaptation sentences.

From the figure we can see that the expressiveness achieved by the adapted HMM sets is closely correlated with the number of sentences used. Furthermore, this expressiveness seems to be affected

(a) (b) (c) (d) (e)

Fig. 5: Example frames from the same sentence, from interpolating the happiness and anger HMM-based AVTTS systems with respective interpolation weights: (a) (0.1, 0.9), (a) (0.3, 0.7), (a) (0.5, 0.5), (a) (0.7, 0.3), and (e) (0.9, 0.1).

by the nature of the emotion adapted each time, since for the emotion of sadness, a high score is obtained with as few as 5 sentences. We believe that this is due to the facts that: (a) the speaking rate of the sadness emotion is slower than the other emotions and closer to the neutral emotion, and (b) that the emotions of happiness and anger are more extreme in terms of facial expressions. In Figure 4 we show example frames from adapting the neutral HMM set to the other three emotions, using 50 adaptation sentences.

4.3. Evaluation of HMM interpolation

To evaluate interpolation of HMM-based AVTTS systems, for each of the 6 emotion pairs that arise from combinations of emotion independent systems, and for each of the weight pairs: (0.9, 0.1), (0.7, 0.3), (0.5, 0.5), (0.3, 0.7), (0.1, 0.9), we interpolated the pairs, and generated 6 unseen sentences. Again via a web-based questionnaire, evaluators were presented with one random sentence from each combination and were asked to recognize the emotion depicted from a list containing the 4 emotions of the corpus. We also included 4 sentences from the emotion-independent systems in order to validate that the emotions are perceived as intended. This questionnaire was completed by 23 respondents. The results of the emotion recognition from the emotion-independent systems are shown in Table 2, where we see that each system achieves a very high classification accuracy.

In Tables 3 and 4 we show the results from interpolating the neutral and sadness HMM sets, and the anger and happiness HMM sets. We can clearly see that we achieve intermediate facial expressions and varying levels of intensity for the emotions, and that they are correlated with the interpolation weights. The abrupt changes in the classification rates also suggest that we need smaller interpolation steps for smoother transitions of the intermediate characteristics. The general consensus is that indeed through HMM interpolation we can achieve varying levels of intensity for emotional audio-visual speech. Furthermore, in Table 4 we can see that the neutral emotion has a high recognition rate, although the interpolation is taking place between anger and happiness. This could suggest that if we gave the respondents more choices on the emotion to recognize, they would pick something else, or that the strength of expressions of anger and happiness is low for these interpolation weights and the confusion causes the viewers to pick the neutral emotion. We also observed the same phenomenon in our results on interpolation in the other emotion pairs. We believe that a further study on this matter is mandatory. Figure 5 shows example frames from interpolating the happiness and anger HMM sets, for our 5 weight pairs.

Table 2: Confusion matrix for the classification of emotions in the emotion-independent HMM systems (% scores).

	Neutral	Anger	Happiness	Sadness
Neutral	90.9	4.55	0	4.55
Anger	4.34	95.65	0	0
Happiness	0	0	100	0
Sadness	0	0	0	100

Table 3: Emotion classification rate when interpolating the neutral and sadness HMM-based AVTTS systems (% scores, w_n: Neutral weight, w_s: Sadness weight).

	Emotions			
(w_n, w_s)	Neutral	Anger	Happiness	Sadness
$(0.1, 0.9)$	8.7	4.35	0	86.96
$(0.3, 0.7)$	18.18	0	0	81.82
$(0.5, 0.5)$	45.83	0	4.17	50
$(0.7, 0.3)$	83.33	0	0	16.67
$(0.9, 0.1)$	91.3	4.35	0	4.35

Table 4: Emotion classification rate when interpolating the anger and happiness HMM-based AVTTS systems (% scores, w_a: Anger Weight, w_h: Happiness Weight).

	Emotions			
(w_a, w_h)	Neutral	Anger	Happiness	Sadness
$(0.1, 0.9)$	0	4.35	95.65	0
$(0.3, 0.7)$	4.35	0	95.65	0
$(0.5, 0.5)$	29.17	16.67	54.17	0
$(0.7, 0.3)$	25	66.67	4.17	4.17
$(0.9, 0.1)$	4.17	95.83	0	0

5. CONCLUSION

In this paper, we tackled the problem of the data overhead that arises when one considers generation of expressive audio-visual speech, with the main focus being the visual modality. To that end, we aimed to equip an AVTTS system with two desired abilities: the ability to adapt to a target emotion with a small amount of adaptation data, and the ability to mix emotions in order generate audio-visual speech with multiple intensity levels and intermediate characteristics.

For the first ability, we employed HMM adaptation in order to adapt a "neutral" HMM-based AVTTS system, where visual modeling is done via an AAM, to the emotions of anger, happiness, and sadness, and showed that we can successfully adapt the expressions of the talking head, even with a small amount of adaptation data. Furthermore, the resulting expressiveness is also correlated with the nature of the target emotion that is being adapted. For the second ability, we employed HMM interpolation between two HMM-based AVTTS systems and showed that we can generate audio-visual speech with different intensity levels for an emotion and with intermediate characteristics between two emotions.

In the future, we aim to investigate other methods of HMM interpolation and adaptation, and for a broader range of emotions, as well as modeling of the AAM parameters through more recent models such as Deep Neural Networks and Recurrent Neural Networks.

6. REFERENCES

[1] J. Bates, "The role of emotion in believable agents," *Communications of the ACM*, vol. 37, pp. 122–125, 1994.

[2] C. Darwin, *The Expression of the Emotions in Man and Animals*, 1871.

[3] A. Mehrabian and S. R. Ferris, "Inference of attitudes from nonverbal communication in two channels.," *J. of consulting psychology*, vol. 31, pp. 248, 1967.

[4] P. Ekman, W. V. Freisen, and S. Ancoli, "Facial signs of emotional experience.," *J. Personality and Social Psychology*, vol. 39, no. 6, pp. 1125, 1980.

[5] R. Plutchik and H. Kellerman, *Emotion, Theory, Research, and Experience: Theory, Research and Experience*, 1980.

[6] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm," *IEEE Trans. Audio, Speech, Language Processing*, vol. 17, pp. 66–83, 2009.

[7] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation for hmm-based speech synthesis system," *Acoustical Science and Technology*, vol. 21, no. 4, pp. 199–206, 2001.

[8] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE Trans. on information and systems*, vol. 88, pp. 2484–2491, 2005.

[9] Z. Wu, S. Zhang, L. Cai, and H. M. Meng, "Real-time synthesis of chinese visual speech and facial expressions using mpeg-4 fap features in a three-dimensional avatar.," in *Proc. Interspeech*, 2006.

[10] X. Li, Z. Wu, H. Meng, J. Jia, X. Lou, and L. Cai, "Expressive speech driven talking avatar synthesis with dblstm using limited amount of emotional bimodal data," in *Proc. Interspeech*, 2016, pp. 1477–1481.

[11] R. Anderson, B. Stenger, V. Wan, and R. Cipolla, "Expressive visual text-to-speech using active appearance models," in *Proc. CVPR*, 2013, pp. 3382–3389.

[12] V. Wan, R. Anderson, A. Blokland, N. Braunschweiler, L. Chen, B. Kolluru, J. Latorre, R. Maia, B. Stenger, K. Yanagisawa, et al., "Photo-realistic expressive text to talking head synthesis.," in *Proc. Interspeech*, 2013, pp. 2667–2669.

[13] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, pp. 681–685, 2001.

[14] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. of Computer Vision*, vol. 60, pp. 135–164, 2004.

[15] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, pp. 33–51, 1975.

[16] G. Papandreou and P. Maragos, "Adaptive and constrained algorithms for inverse compositional active appearance model fitting," in *Proc. CVPR*, 2008, pp. 1–8.

[17] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The hmm-based speech synthesis system (hts) version 2.0.," in *Proc. ISCA SSW6*, 2007, pp. 294–299.

[18] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-markov model-based speech synthesis system," *IEICE Trans. on information and systems*, vol. 90, pp. 825–834, 2007.

[19] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," in *Proc. Eurospeech*, 1999, pp. 2347–2350.

[20] J. J. Odell, *The Use of Context in Large Vocabulary Speech Recognition*, Ph.D. thesis, Univesity of Cambridge, 1995.

[21] K. Shinoda and T. Watanabe, "Acoustic modelling based on the mdl principle for speech recognition," in *Proc. EUROSPEECH*, 1997, pp. 99–102.

[22] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *Proc. ICASSP*, 2000, vol. 3, pp. 1315–1318.

[23] V. V. Digalakis and L. G. Neumeyer, "Speaker adaptation using combined transformation and bayesian methods," *IEEE Trans. Speech and Audio Processing*, vol. 4, pp. 294–300, 1996.

[24] J. Yamagishi, T. Masuko, and T. Kobayashi, "Hmm-based expressive speech synthesis - towards tts with arbitrary speaking styles and emotions," in *Proc. of Special Workshop in Maui*, 2004.

[25] A. Katsamanis, M. Black, P. G. Georgiou, L. Goldstein, and S. Narayanan, "Sailalign: Robust long speech-text alignment," in *Proc. VLSP*, 2011.

[26] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *Proc. MAVEBA*, 2001, pp. 59–64.