Speech Communication 95 (2017) 137–152

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Video-realistic expressive audio-visual speech synthesis for the Greek language

Panagiotis Paraskevas Filntisis [a,b,*], Athanasios Katsamanis [a,b], Pirros Tsiakoulis [c], Petros Maragos [a,b]

[a] School of Electrical and Computer Engineering, National Technical University of Athens, Zografou Campus, Athens, 15773, Greece
[b] Athena Research and Innovation Center, Maroussi, 15125, Greece
[c] INNOETICS LTD, Athens, Greece

ARTICLE INFO

Article history:
Received 22 January 2017
Revised 5 June 2017
Accepted 28 August 2017
Available online 31 August 2017

Keywords:
Audio-visual speech synthesis
Expressive
Hidden Markov models
Deep neural networks
Interpolation
Adaptation

ABSTRACT

High quality expressive speech synthesis has been a long-standing goal towards natural human-computer interaction. Generating a talking head which is both realistic and expressive appears to be a considerable challenge, due to both the high complexity in the acoustic and visual streams and the large non-discrete number of emotional states we would like the talking head to be able to express. In order to cover all the desired emotions, a significant amount of data is required, which poses an additional time-consuming data collection challenge. In this paper we attempt to address the aforementioned problems in an audio-visual context. Towards this goal, we propose two deep neural network (DNN) architectures for Video-realistic Expressive Audio-Visual Text-To-Speech synthesis (EAVTTS) and evaluate them by comparing them directly both to traditional hidden Markov model (HMM) based EAVTTS, as well as a concatenative unit selection EAVTTS approach, both on the realism and the expressiveness of the generated talking head. Next, we investigate adaptation and interpolation techniques to address the problem of covering the large emotional space. We use HMM interpolation in order to generate different levels of intensity for an emotion, as well as investigate whether it is possible to generate speech with intermediate speaking styles between two emotions. In addition, we employ HMM adaptation to adapt an HMM-based system to another emotion using only a limited amount of adaptation data from the target emotion. We performed an extensive experimental evaluation on a medium sized audio-visual corpus covering three emotions, namely anger, sadness and happiness, as well as neutral reading style. Our results show that DNN-based models outperform HMMs and unit selection on both the realism and expressiveness of the generated talking heads, while in terms of adaptation we can successfully adapt an audio-visual HMM set trained on a neutral speaking style database to a target emotion. Finally, we show that HMM interpolation can indeed generate different levels of intensity for EAVTTS by interpolating an emotion with the neutral reading style, as well as in some cases, generate audio-visual speech with intermediate expressions between two emotions.

1. Introduction

Human-computer interaction (HCI) has exploded in the latest decades and, virtual or physical, intelligent agents[1] have become an integral part of a person's everyday life. These agents take part in a variety of applications, either of small significance such as everyday human tasks, or of great significance such as teaching assistants for pedagogical purposes (Johnson et al., 2000). Artificial Intelligence (AI) aims to maximize the naturalness of human-computer interactions so that the human forgets that he is interacting with a computer. Taking into account the fact that the main mode of human communication is speech, we understand that speech synthesis constitutes a vital part of AI for human-computer communication. Speech synthesis does not include only the generation of a human-like voice; speech is multimodal in nature (McGurk and MacDonald, 1976; Ekman, 1984) and important

* Corresponding author.
E-mail addresses: filby@central.ntua.gr (P.P. Filntisis), nkatsam@cs.ntua.gr (A. Katsamanis), ptsiak@innoetics.com (P. Tsiakoulis), maragos@cs.ntua.gr (P. Maragos).

[1] An agent is anything that can be viewed as perceiving its environment through sensors, and acting upon that environment through effectors (Russell and Norvig, 1995). Example of a physical agent is a robot, while of a virtual agent is an avatar.

http://dx.doi.org/10.1016/j.specom.2017.08.011
0167-6393/

information is included in the visual stream of information (i.e., the human face, and more generally the human head and its movements) along with the acoustic stream. It has been shown that the inclusion of a visual stream of information increases the intelligibility of speech, especially under noise, even when the face is a virtual talking head (Sumby and Pollack, 1954; Ouni et al., 2007).

Achieving a high degree of naturalness in HCI is highly correlated with the ability of the agent to express emotions. Agents capable of expressiveness are more believable and life-like, thus have a stronger appeal to their interlocutor (Bates, 1994). In addition, expressive behavior itself contains important information (Ambady and Rosenthal, 1992) and affects the emotional state of the other party (Hatfield et al., 1993; Keltner and Haidt, 1999) and, consequently, its decision making (Schwarz, 2000; Bechara, 2004).

Strongly correlated with speech, emotion is conveyed multimodally (Darwin, 1871; Richard et al., 2002), so the agent must be capable of expressing emotion multimodally as well. It is also reasonable to assume that emotions should be expressed through all the information channels simultaneously; otherwise it is possible that the receptor of the signals might become confused in regards to the emotional state of the agent, since neurological studies have shown that the perception of the acoustic and the visual streams affect each other (Skipper et al., 2007). Under the same assumption it is also desirable that the levels of expressiveness of both information channels are correlated - i.e., a full blown facial expression of anger is accompanied with a full blown vocal expression of anger.

Audio-Visual Text-To-Speech synthesis (AVTTS) explores the generation of audio-visual speech signals (Mattheyses and Verhelst, 2015) (i.e., a talking head), and video-realistic AVTTS, more specifically, explores the generation of talking heads that highly resemble a human being as if a camera was recording it. Although naturalness in video-realistic AVTTS systems has increased greatly, the addition of expressions has proven to be a challenging task (Anderson et al., 2013; Schröder, 2009) due to the large variability they introduce, especially in extreme expressions such as expressions of anger or happiness, in both acoustic and visual modeling. This large complexity increases the probability of introducing artifacts in the generated face and causing the “uncanny valley” effect (Seyama and Nagayama, 2007; Mori et al., 2012).

Motivated by the above, in this paper we study and improve video-realistic expressive audio-visual text-to-speech synthesis (EAVTTS), by tackling several challenges that arise when considering the addition of expressiveness.

When considering synthesis of expressive speech we would ideally desire that the agent is able to express itself in the same ways a human can, in order to achieve the maximum level of naturalness. Studies on the nature of emotions have claimed both that some emotions can be defined as the combinations of others (Plutchik, 2001; Plutchik and Kellerman, 1980; Plutchik, 1980), and that each emotion has different levels of intensity which are expressed with variations between each other (Ekman, 1984; Ekman et al., 1980; Hess et al., 1997). We believe that these studies can be proven useful when considering the problem of the number of emotional states the agent must be able to express. If we consider that the expression of emotions can also be defined as combinations of other emotions, we could instruct the agent to express more complex expressions, and speech with intermediate style between different emotions, through mixtures of emotions that it “knows” how to perform. In the same manner, considering that expressions vary with the intensity of the emotion, if we see the different intensity levels of an emotion as a combination of the emotion with the neutral expression/voice, we can express emotions in different continuous intensity levels.

Closely correlated with the number of emotional states and intensity levels we would like to be able to synthesize, is the ability of an EAVTTS system to adapt to a target emotion, given some examples of it.

Considering the two desired abilities we just stated, we argue that Unit Selection (US) synthesis is not suitable for extensions to expressive speech. US synthesis is more natural than parametric synthesis (under the assumption of a large enough training set) because the imperfect reconstruction of speech from parameters is avoided, however, immense data needs to be collected for variations in speaking style (Black, 2003). Parametric synthesis on the other hand, appears to be suitable for the task in hand due to its flexibility that arises from the statistical modeling process (Zen et al., 2009), which allows modification of voice characteristics and speaking style. The same conclusion holds for EAVTTS as well; US suffers from low flexibility in changing facial expressions, while a parametric model allows us to do so.

Taking into account our previous introduction, in this paper we do a much needed thorough exploration of video-realistic AVTTS and propose methods to tackle the previously stated challenges and achieve the desired EAVTTS abilities: high complexity, continuous and complex emotional states, and adaptiveness.

Motivated by the recent advances in speech synthesis with deep learning (Ling et al., 2015), we propose two different deep neural network (DNN) pipelines for EAVTTS and examine the level at which these architectures are able to model the special characteristics and the full extent of expressive speech in an audio-visual context. This examination is achieved through direct comparison with the traditional parametric approach of HMM-based EAVTTS, as well as a concatenative unit selection EAVTTS system, both on the realism as well as on the expressiveness of the generated talking head, when the systems are trained on a corpus featuring expressive audio-visual speech.

As far as HMM-based EAVTTS is concerned, we also assess the level of emotional strength acquired by an HMM-based AVTTS system that has been adapted to another emotion, by directly comparing with an HMM-based AVTTS system that has been trained on the full corpus of the respective emotion. We also employ HMM model interpolation in order to generate audio-visual speech with different intensity levels of expressiveness and more complexity. To our knowledge, there has not yet been a study to show the results of HMM adaptation and interpolation for HMM-based AVTTS.

Our final contribution is a new medium-sized corpus featuring expressive audio-visual speech in three emotions: anger, happiness, and sadness, plus neutral reading style (which we will refer to as the neutral emotion from now on), from one speaker and for the Greek language.

The remainder of the paper is organized as follows: In the next section we present the background and related works on parametric video-realistic AVTTS and EAVTTS. In Section 3 we present the proposed DNN-based expressive audio-visual speech synthesizers and in Section 4 we present a traditional HMM-based expressive audio-visual speech synthesis system along with the methods of adaptation and interpolation we employ. Next, we briefly describe the unit selection EAVTTS system we built for comparing with our parametric approaches and in Section 6 we present the CVSP-EAV (CVSP Expressive Audio-Visual Speech) corpus, which we collected for evaluating our approaches. Section 7 presents our experimental results and the last section our conclusive remarks and future work directions.

2. Background/related work

Audio-visual speech synthesis (AVTTS) can be divided into two distinguishable categories based on the way the talking head is synthesized (Bailly et al., 2003). The first category includes 2D/3D graphical models typically constructed by a mesh of polygons and vertices, and the creation of facial expressions involves the move-

(a) Cao et al. (2005) (b) Ezzat et al. (2002) (c) Anderson et al. (2013) (d) Xie et al. (2014) (e) Fan et al. (2015)

Fig. 1. Various examples of video-realistic talking heads.

ment of the mesh. Examples of model-based visual or audio-visual speech synthesis include (Schabus et al., 2014; Salvi et al., 2009; Sifakis et al., 2006; Beskow, 1996; Le Goff and Benoît, 1996; Pelachaud et al., 1996).

The second category of image-based AVTTS is driven by a training set of image sequences. Based on frontal recordings of a speaking person, snapshots of the person speaking arbitrary utterances not found in the recordings are generated. This category of image-based or video-realistic AVTTS can be further divided in unit selection and parametric methods.

Unit selection methods typically use raw or slightly modified images (and waveforms for audio) in the training set in order to construct the target audio-visual sequence (Cosatto et al., 2000; Mattheyses et al., 2008; Huang et al., 2002). In these architectures, typically the cost that is minimized by the unit selection module involves both acoustic and visual features.

Parametric methods involve the training of a statistical model controlled by a small number of parameters that can be used to reconstruct photo-realistic frames. Examples of parametric visual models in this category involve Active Appearance Models (Cootes et al., 2001), Eigenfaces (Turk and Pentland, 1991) and Multidimensional Morphable Models (MMMs) (Jones and Poggio, 1998).

A first example of parametric video-realistic audio-visual speech synthesis which is based on HMMs and uses the same pipeline as HMM-based text-to-speech synthesis (TTS) (Zen et al., 2007b) is in Tamura et al. (1999) and Sako et al. (2000), where human lips recordings are modeled through a technique similar to eigenfaces (eigenlips in this case), and the eigenlips weights are added to the observation vector along with the acoustic parameters (mel-cepstral coefficients and fundamental frequency). Ezzat et al. (2002) employs MMMs and combines them with a custom trajectory synthesis technique for generating video-realistic speech. In Xie et al. (2014) and Xie and Liu (2007), a lower-face Active Appearance Model combined with HMMs (and their variations) is used to generate full-face audio-visual speech by employing Poisson image stitching (Pérez et al., 2003). Fan et al. (2015) also used a similar technique for full-face video-realistic generation and successfully employed a Bidirectional Long Short Term Memory network in order to predict the weights of a lower-face Active Appearance Model, from a small number of linguistic features. Hybrid methods where the selection of images from the corpus is driven by HMM modeling have also been proposed (Wang et al., 2010; Mattheyses et al., 2011).

In the field of expressive AVTTS (EAVTTS) with 2D/3D models, some direct approaches have included the modeling of human facial expressions with a set of parameters (most commonly FAPs - facial animation parameters (Pandzic and Forchheimer, 2003)), and then using these parameters to drive a graphics based 3D model (Wu et al., 2006; Deng et al., 2006; Li et al., 2016). Image-based unit selection methods for generating video-realistic expressive speech have been presented in Cao et al. (2005), Melenchón et al. (2009) and Liu and Ostermann (2011).

Parametric video-realistic EAVTTS has not seen many studies. An example of an expressive video-realistic talking head using Active Appearance Models and HMM modeling was presented in Anderson et al. (2013) and Wan et al. (2013), where AAM modeling of the face was also extended to alleviate local facial deformations such as blinking, and remove the facial pose. This system used cluster adaptive training of HMMs in order to model expressions of different emotions, as well as generate combinations of emotions. In Shaw and Theobald (2016), modeling of emotional expressions is achieved using AAMs and Independent Component Analysis (ICA). Furthermore, ICA is employed to generate mixtures of expressions. Fig. 1 shows various examples of video-realistic talking heads from previous works.

While studies on parametric video-realistic EAVTTS and manipulation of expressions are scarce, tools are available to adapt a trained HMM-based acoustic speech synthesis system to a new speaking style or speaker, using a small amount of adaptation data. Such adaptation methods include maximum-likelihood linear regression (MLLR) (Leggetter and Woodland, 1995; Tamura et al., 2001), maximum a posteriori adaptation (MAP) (Digalakis and Neumeyer, 1996; Masuko et al., 1997), and their variations (e.g., constrained MLLR (CMLLR) (Gales, 1998) and Constrained Structural Maximum a Posteriori Linear Regression (CSMAPLR) (Yamagishi et al., 2009)). HMM adaptation has already been employed successfully in Yamagishi et al. (2004) for adapting an acoustic neutral HMM-based system to the emotions of joy and sadness. However, the results of HMM adaptation when considering HMM-based AVTTS have not yet been studied.

Similarly, HMM interpolation has been proposed for HMM-based acoustic speech synthesis, by building different systems for each different speaking style/speaker and then interpolating the Gaussian output distributions of the models using arbitrary weights (Yoshimura et al., 2000; Yamagishi et al., 2004; Tachibana et al., 2005). As with HMM adaptation, HMM interpolation has yet to be applied to AVTTS, but it has successfully been employed again in Yamagishi et al. (2004) for interpolation between HMM sets trained on different emotions.

3. DNN-based expressive audio-visual speech synthesis

In our two proposed DNN-based architectures for expressive audio-visual speech synthesis, each emotion is modeled separately, by a different DNN-based audio-visual synthesizer (we will use the term EAVTTS system when considering the full system that models all emotions, while the term AVTTS system denotes the subsystems). These sub-synthesizers (or subsystems) follow one of the two architectures that can be seen in Figs. 2 and 3. Acoustic, visual, and linguistic features are extracted from an audio-visual corpus and then used in order to train the neural network subsystems of each architecture. We will first describe the features that we employ for audio-visual modeling, and then describe the two different DNN AVTTS models that are the components of the two different DNN-based EAVTTS architectures.

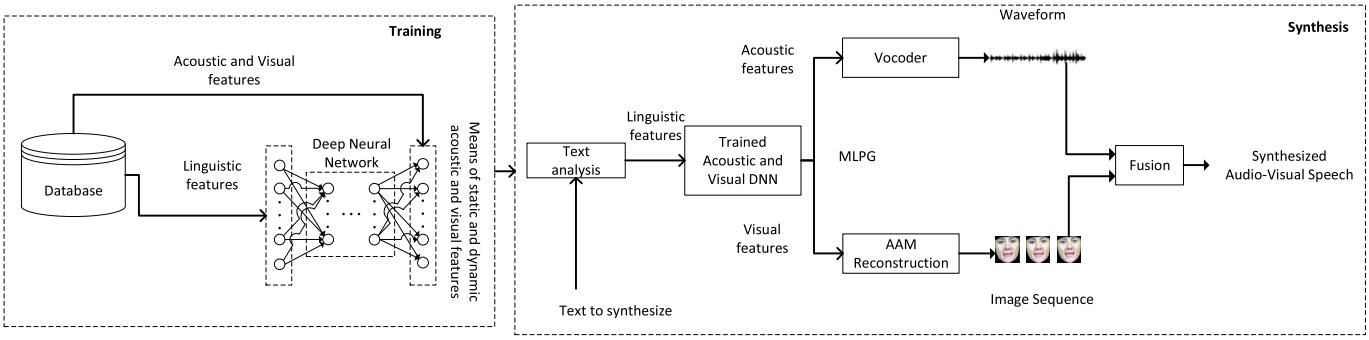

Fig. 2. DNN-based audio-visual speech synthesis with joint modeling of acoustic and visual features.

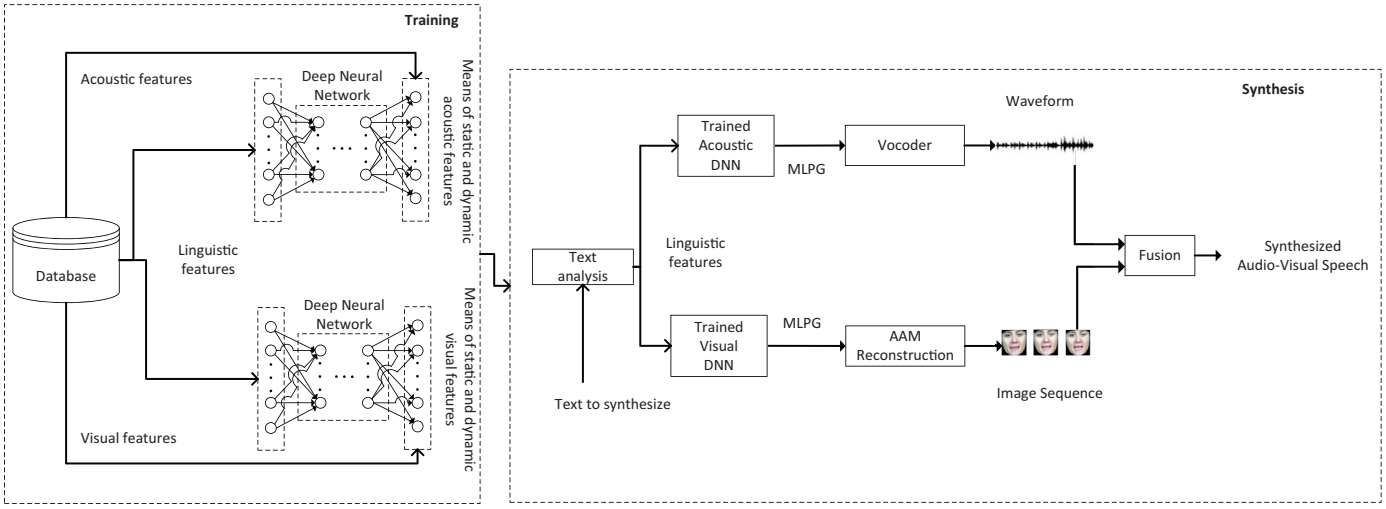

Fig. 3. DNN-based audio-visual speech synthesis with separate modeling of acoustic and visual features.

3.1. Features for DNN-based audio-visual speech synthesis

Acoustic Features. Speech is modeled by mel-frequency cepstral coefficients (MFCCs), the logarithmic fundamental frequency, and band-aperiodicity coefficients using STRAIGHT analysis (Kawahara et al., 1999, 2001). To reconstruct the waveform from the spectral and prosodic features, the STRAIGHT vocoder is used.

Visual Features. To obtain a low dimensional parametric model of the face for all of the different emotions, we employ an Active Appearance Model (AAM) (Cootes et al., 2001; Matthews and Baker, 2004). We model the whole face and not only the lower part since emotional expressions include the upper facial half as well.

In active appearance modeling, a face (and more generally the object modeled) consists of the shape and the texture. The shape is represented by a vector $\boldsymbol{s}$, the elements of which are the coordinates of M vertices that make up the mesh of the face. For a particular snapshot (frame), the shape is expressed as the mean shape $\bar{\boldsymbol{s}}$ (that is the mean of the coordinates of the vertices of several frames after a Procrustes analysis is applied to them), plus a linear combination of n eigenvectors (called eigenshapes) $\boldsymbol{s}_i$ that are found via employing a Principal Component Analysis (PCA) to the training meshes:

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

where p_i is the weight applied to the eigenshape $\boldsymbol{s}_i$.

The texture of the face is modeled in the same way as the shape, after normalizing the shape of each training mesh using an affine transformation or another method such as thin plate splines:

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

where $\boldsymbol{A}(\boldsymbol{x})$ is texture defined over the pixels $\boldsymbol{x}$ that lie in the mesh of the mean shape $\bar{\boldsymbol{s}}$, $\bar{\boldsymbol{A}}(\boldsymbol{x})$ is the mean texture, $\boldsymbol{A}_i(\boldsymbol{x})$ are the eigenvectors found via PCA (called eigentextures) and λ_i is the weight applied to the eigentexture $\boldsymbol{A}_i(\boldsymbol{x})$.

Upon obtaining the weights of the eigenshapes and the eigenvectors of a snapshot of the face, the image can be reconstructed by warping the texture $\boldsymbol{A}(\boldsymbol{x})$ from the mean shape $\boldsymbol{s}_0$ to the computed shape $\boldsymbol{s}$ based on a warp $\boldsymbol{W}(\boldsymbol{x}; \boldsymbol{p})$, where $\boldsymbol{p}$ is the vector of the shape weights p_i.

During the fitting of an active appearance model to a new frame of the modeled object, if we denote as $\boldsymbol{I}(\boldsymbol{x})$ the texture of the frame defined over the pixels $\boldsymbol{x}$, we seek to minimize the euclidean norm (called the reconstruction error):

$$E = \|I(\boldsymbol{W}(\boldsymbol{x}; \boldsymbol{p})) - \boldsymbol{A}(\boldsymbol{x})\|_2^2 \tag{3}$$

where $\boldsymbol{I}(\boldsymbol{W}(\boldsymbol{x}; \boldsymbol{p}))$ is the warped back image texture and $\boldsymbol{A}(\boldsymbol{x})$ is the synthesized texture.

Building an AAM for a large database depicting different expressions (and extreme ones that arise during emotions such as happiness or anger) appears to be a challenging task. Due to the large variations introduced, minimization of the reconstruction error usually results in undesirable results. For this reason, we incorporate prior constraints (Papandreou and Maragos, 2008) during the fitting, as a means of increasing the robustness of the model.

In a model including prior constraints, the error minimized is:

$$E_p = \|\boldsymbol{I}(\boldsymbol{W}(\boldsymbol{x};\boldsymbol{p})) - \boldsymbol{A}(\boldsymbol{x})\|_2^2 + Q(\boldsymbol{q}) \quad (4)$$

where $Q(\boldsymbol{q})$ is a quadratic penalty corresponding to a prior Gaussian distribution with mean $\boldsymbol{q}$. More information on the prior constraints can be found in Papandreou and Maragos (2008).

3.2. Architectures for DNN-based audio-visual speech synthesis

The two DNN-based AVTTS synthesizers of each EAVTTS architecture differ in the fact that in Fig. 2, the neural network maps linguistic features to acoustic and visual features at the same time, whilst in Fig. 3, this mapping is done separately for the acoustic and visual features by two different neural networks. The linguistic features contain information about the lexicological context of the current phoneme and can consist of either answers to binary questions (e.g., "is the current phoneme a vowel?") or numerical values (e.g., the number of syllables in a word). Within-phone positional features such as position of the acoustic/visual frame within the state of the phone (in an HMM context), phone and state duration, and state position within the phone (Zen et al., 2013; Wu et al., 2016) are also included. The output (acoustic, visual or joint audio-visual) features also include dynamic features (first and second derivatives of static features).

In both AVTTS synthesizers, a neural network (not shown in the figures) is employed for predicting the duration of speech. In the network responsible for duration modeling, an input vector containing frame-level linguistic features is mapped to durations of either the phoneme considered, or the phoneme states (in an HMM context).

During the training phase, linguistic features extracted from the database, along with acoustic and visual features, are used to train the networks. The mapping from linguistic features to acoustic, visual or joint audio-visual features constitutes a regression problem, and the following mean squared error function is minimized by the network with a Backpropagation procedure (Williams and Hinton, 1986):

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2 \quad (5)$$

where N is the number of output features, $\hat{y}_i$ is the ith predicted feature and y_i is the real outcome.

In the synthesis stage, after analyzing the text to be synthesized and extracting its linguistic feature representation, the neural networks predict the output acoustic, visual, or joint audio-visual features. The outputs of the neural networks are considered to be the mean vector of the acoustic and visual features while the covariances are pre-calculated from the training data. The Maximum-Likelihood Parameter Generation algorithm (Tokuda et al., 2000) is then used in order to produce smooth trajectories of acoustic and visual features. This step is imperative in order to alleviate the fact that DNNs do not have memory or take into account adjacent frames during training (Zen, 2015). Postfiltering in the cepstral domain (Yoshimura et al., 2005) is applied to acoustic features.

In general, DNN-based synthesis (both audio and audio-visual) possesses important advantageous properties as opposed to HMM-based synthesis (Zen et al., 2013; Watts et al., 2016; Qian et al., 2014):

1. Deep layered architectures can represent highly complex function transformations compactly.
2. In DNN-based AVTTS, contrary to HMM-based AVTTS where predictions take place on a state level, predictions take place on an acoustic frame level.
3. Decision trees are incapable of modeling complex dependencies between input features whereas DNNs can compactly model these dependences. Furthermore, decision trees perform a hard split of the linguistic space which results in inferior generalization to DNN-based modeling where the weights of the network are trained using the whole training set.
4. Linguistic features can also hold numerical and not only binary values. Zen et al. (2013) found out in experimental results that numerical values perform better and more efficiently.

4. HMM-based expressive audio-visual speech synthesis

Similarly with our proposed DNN-based EAVTTS systems, the HMM-based EAVTTS architecture models each emotion separately, through multiple HMM AVTTS systems. In this section we will do an overview of HMM-based audio-visual speech synthesis and then describe the methods of adaptation and interpolation that the system employs in order to adapt to new emotions, and mix known emotions.

4.1. Overview of HMM-based audio-visual speech synthesis

The architecture of an HMM-based audio-visual speech synthesis system is shown in Fig. 4. Typically, the same pipeline with HMM-based acoustic speech synthesis (Tokuda et al., 2013) is employed.

Multi-Space probability Distribution Hidden Semi Markov Models (MSD-HSMMs) (Zen et al., 2007b; Heiga et al., 2007; Yoshimura et al., 1999) are used to model both the acoustic and visual features of speech simultaneously, so as to enforce a strong temporal alignment of the visual and acoustic streams.

The observation vector also contains dynamic features, in order to avoid discontinuities that arise from the step-wise sequence that is generated in the synthesis stage.

To alleviate the problem of limited training data, a decision tree based context clustering method is applied (Odell, 1995). During the decision tree clustering approach, using a predefined set of contextual questions, each node is split into two, by choosing the question that minimizes the Description Length (Shinoda and Watanabe, 1997) of the data. Upon termination, states that belong in the same terminal node (leaf) of the tree are merged.

In the synthesis part, the maximum-likelihood parameter generation algorithm (Tokuda et al., 2000) is used to generate smooth trajectories of both the acoustic and visual parameters from the static and dynamic parameters emitted from the HSMMs. Just like in the DNN architectures, postfiltering in the cepstral domain (Yoshimura et al., 2005) is applied to acoustic features.

4.2. Adaptation for HMM-based EAVTTS

In order to tackle the data collection overhead that arises when considering expressive audio-visual speech, we use HMM adaptation to adapt an audio-visual HMM set that is trained on a neutral training set, to a target emotion, using adaptation data that depict the emotion. The algorithm we employ is the CSMAPLR (Constrained Structural Maximum a Posteriori Linear Regression) adaptation method (Yamagishi et al., 2009, 2007).

The CSMAPLR adaptation method combines the advantages of both SMAP (Shinoda and Lee, 2001) and CMLLR (Gales, 1998) methods, and makes use of linguistic information, through the regression tree that is used to propagate the prior information from the root of the tree, to the lower nodes. In addition, because the method is employing recursive MAP estimation (Digalakis and Neumeyer, 1996), it is robust when a low amount of adaptation data is available (Lorenzo-Trueba et al., 2015). After the CSMAPLR adaptation, an additional MAP adaptation is applied.

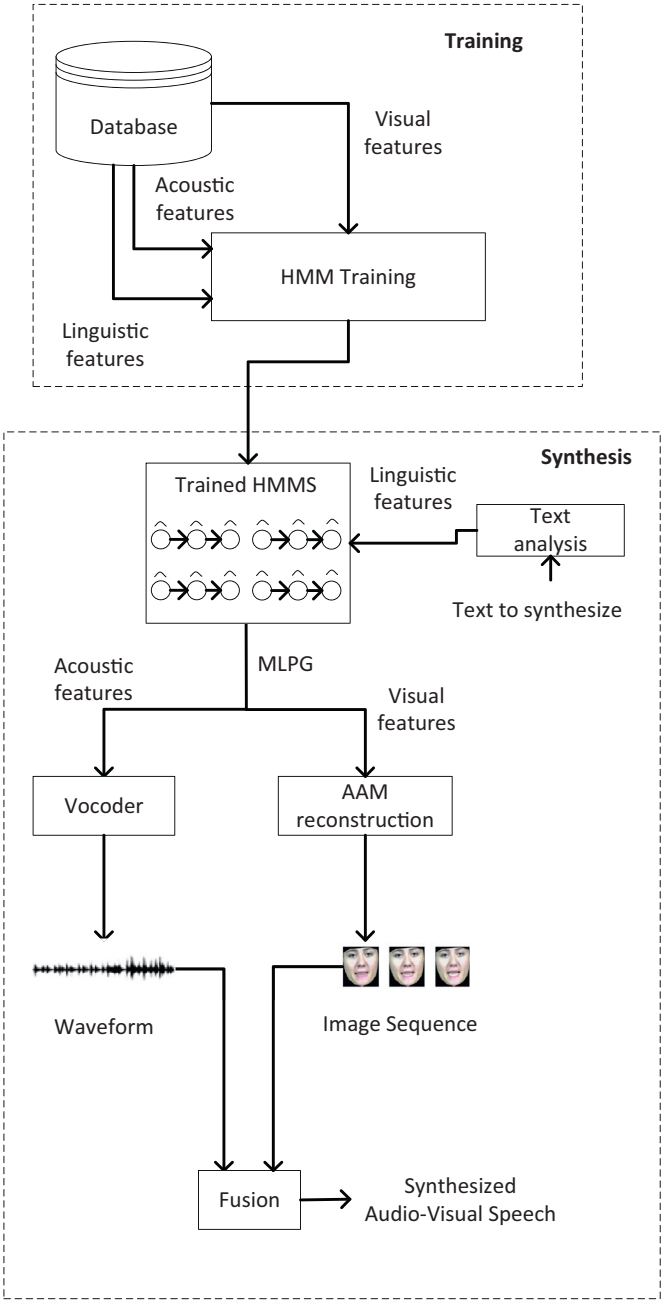

Fig. 4. HMM-based audio-visual speech synthesis system architecture.

In the CSMAPLR adaptation, the mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of a Gaussian state-output or state-duration distribution, are transformed simultaneously through the transformation matrix $\boldsymbol{Z}$ and the transformation bias $\boldsymbol{\epsilon}$:

$$\bar{\boldsymbol{\mu}} = \boldsymbol{Z}\,\boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{6}$$

$$\bar{\boldsymbol{\Sigma}} = \boldsymbol{Z}\,\boldsymbol{\Sigma}\,\boldsymbol{Z}^{\top} \tag{7}$$

4.3. Interpolation for HMM-based EAVTTS

Interpolation between emotions not only allows us to obtain emotions with various intensity levels and form more complex speaking styles and expressions, but also offers us the ability to control more formally the resulting expressions through the use of the interpolation weights, as opposed to adaptation methods.

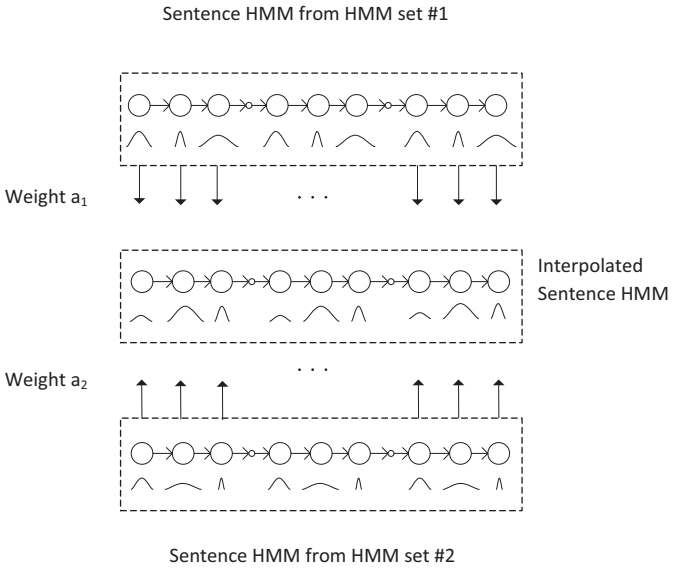

Fig. 5. Interpolation for HMM-based EAVTTS. Figure adapted from Tachibana et al. (2005).

In Yoshimura et al. (2000) interpolation of HMM models trained on different datasets is done via maximization of the KL divergence between the output Gaussian distributions of the models. Yamagishi et al. (2004) follow a simpler approach (interpolation between observations) for the interpolation of the HMM models, which we adopt as well for the interpolation of HMM sets trained on our four emotional training sets.

If we denote the output Gaussian distribution of an HMM state as $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is the mean vector and covariance matrix of the distribution respectively, it becomes apparent that for HMM-based speech synthesis, the problem of interpolating emotions corresponds to interpolation of Gaussian distributions for respective HMM states across systems trained on a different training set which represent a speaking style.

The mean and variance of the interpolated pdf $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is:

$$\boldsymbol{\mu} = \sum_{i=1}^{K} \alpha_i \boldsymbol{\mu}_i \tag{8}$$

$$\boldsymbol{\Sigma} = \sum_{i=1}^{K} \alpha_i^2 \boldsymbol{\Sigma}_i \tag{9}$$

where K is the number of the different pdfs that will be interpolated and a_i is the weight corresponding to the ith pdf. This interpolation is applied to the duration models as well. It is noted that the weights are chosen so that:

$$\sum_{i=1}^{K} |\alpha_i| = 1 \tag{10}$$

To deal with the fact that HMM sets trained on different datasets have a different tying structure, the interpolation of the pdfs is done on the synthesis level, after constructing a sentence HMM from each HMM set for the specific label to be synthesized. Fig. 5 depicts this approach.

5. Unit selection audio-visual speech synthesis

In this section, we describe the unit selection video-realistic EAVTTS system, which employs multiple US AVTTS subsystems to model each different emotion. The subsystems are based on the unit selection acoustic speech synthesis system described in

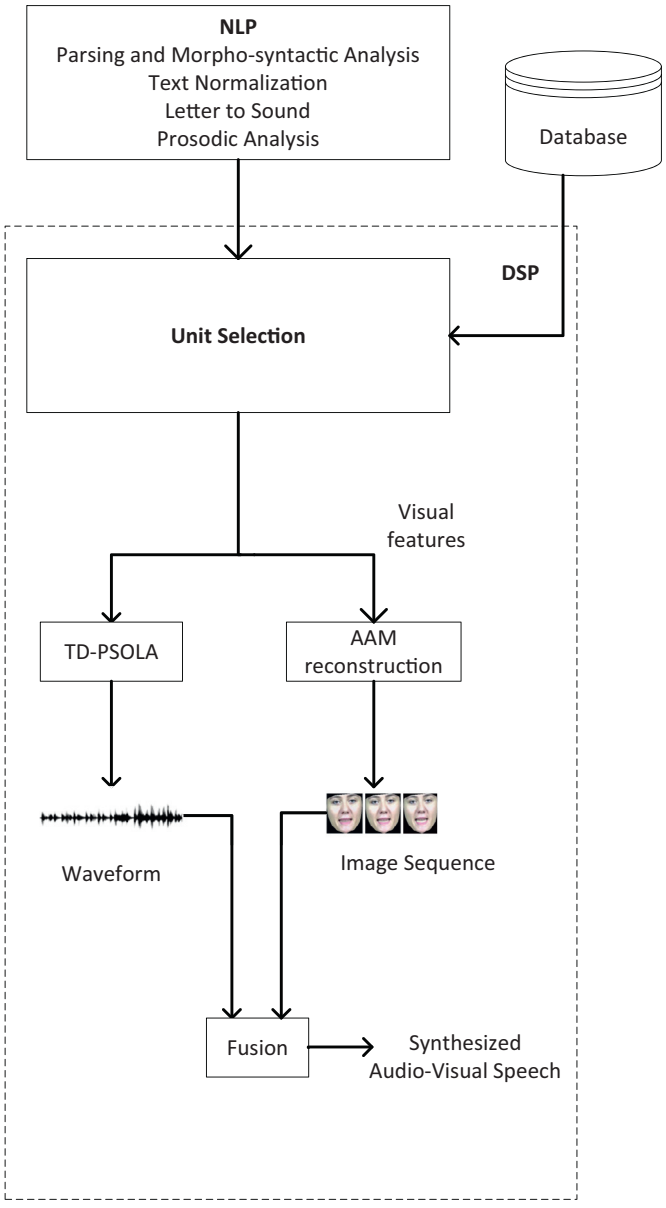

Fig. 6. Unit selection based audio-visual speech synthesis.

Raptis et al. (2016) and Chalamandaris et al. (2013) and were modified to include the visual modality as well. We did that in order to have a direct comparison of a concatenative EAVTTS system against our parametric systems. The system utilizes its own front end (as opposed to our previous methods) and its architecture is shown in Fig. 6.

The subsystems follow a typical concatenative unit selection architecture, split into two components:

The NLP (natural language processing) component which is responsible for extracting all relevant information from the input text and transforming it into an intermediate format. This component comprises of the following modules: a word- and sentence- tokenization module, a text normalizer, a letter-to-sound module and a prosody generator.

The DSP (digital signal processing) component consists of the unit selection module, the signal manipulation module that generates the speech waveform, and the image reconstruction module, which is essentially the same module used in the previously described methods that reconstructs the image sequence and joins it with the speech waveform.

The unit selection module of the system optimizes a cost function that consists of two terms: the target cost, which is the cost of similarity of phonetic and prosodic context between two units, and the join cost, affected by pitch continuity, spectral similarity, and visual similarity. The visual similarity is integrated into the unit selection cost function as two additional terms in the join cost function; one for each of the visual feature vectors, namely the shape and texture feature vectors. The Euclidean norm is used as the distance between the shape and texture vectors. The modified join cost function is a weighted sum of the auditory components, i.e. the pitch and spectral cost functions, and the visual components, i.e. shape and texture. The weights are chosen so that all components have equal range, i.e. assigning equal importance to both modalities.[2]

The final waveform of speech is generated using a custom Time Domain Pitch Asynchronous Overlap Add (TD-PSOLA) method to concatenate the units selected by the unit selection module.

The final image sequence is generated, by concatenating the visual parameters (eigenshape and eigentexture weights) that correspond to the audio-visual units selected by the unit selection module. During this concatenation we do not employ a smoothing technique.

6. The CVSP-Expressive Audio-Visual Speech Corpus

6.1. Recording of the corpus

In order to evaluate the methods described in this study, we collected a medium sized corpus featuring expressive audio-visual speech in Greek for 4 emotions (neutral, anger, sadness and happiness). The database, which we call CVSP-EAV (CVSP - Expressive Audio-Visual) Corpus, was recorded in an anechoic studio at the Athena Innovation Research Center. A professional actress was hired to act the aforementioned emotions. The actress was instructed to express the emotions in an extreme and clear manner. Although we are aware that humans rarely feel or express emotions in an extreme manner - and as such a talking head expressing extreme emotions would feel unrealistic in most HCI cases - since we are exploring modeling of emotional speech, it is reasonable to take into account the extreme case of each specific emotion. Furthermore, this allows us to correctly evaluate the synthesis of different intensity levels corresponding to each emotion which can range from reading style (neutral) to the full-blown expression of the emotion.

The actress was seated in from of a high-definition camera recording video in 1080p resolution at 29.97 frames per second in a H.264 format. A high quality microphone was used to capture the audio with a sampling rate of 44,100 Hz.

The textual corpus consists of 900 sentences in Greek which were selected so that the corpus would have a phonetic distribution representative of the Greek language. Each sentence was pronounced by the actress 4 times, once for each of the four emotions: anger, sadness, happiness, and neutral, resulting in a total of 3600 sentences.

Fig. 7 depicts a sample image from the recordings for each different emotion.

6.2. Processing of the corpus

Because the recording of all the sentences in the corpus was continuous, in order to split the recorded video and audio in

[2] Fine-tuning the unit selection weights requires extensive listening/visual experiments which is outside the scope of this work.

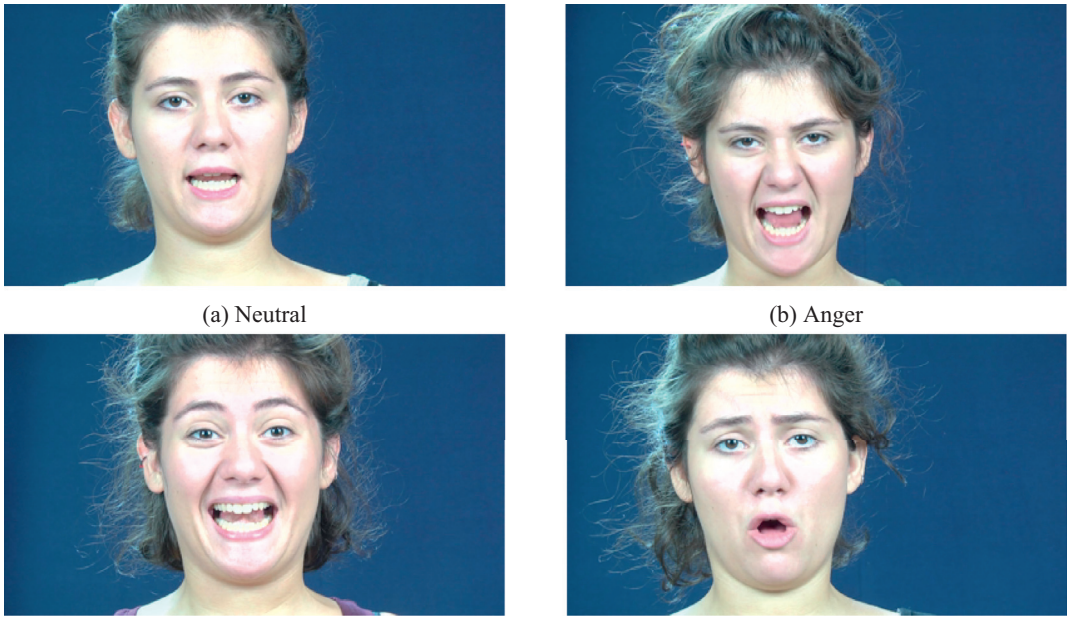

Fig. 7. Sample images for each of the different emotions present in the CVSP-EAV Corpus.

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

	Neutral	Anger	Happiness	Sadness
Sentences	899	898	896	894
Duration (minutes)	72	71	72	86
Frames (approx.)	129,000	129,000	129,000	150,000

the different sentences the actress pronounced, we employed the sail-align toolkit (Katsamanis et al., 2011) which contains hidden Markov models (HMM) for the Greek language, trained in the Logotypografia database (Digalakis et al., 2003), featuring ~72 h of speech from 125 speakers. The toolkit employs a three step speaker adaptation algorithm in order to increase the accuracy of the alignments. This alignment method was applied to the full recording of each of our four different emotions, and apart from obtaining the splits of the recordings at a sentence level, we also obtained four different speaker dependent HMM sets adapted to each of the different emotions in the corpus.

Next, the audio recorded from the high quality microphone was temporally aligned with the video using the cross-correlation between the high quality sound and the sound from the camera, and the frames from the video corresponding to each sentence were extracted in high quality JPEG format.

Finally, we force aligned at the phoneme level each sentence with its transcription, using the previously obtained adapted models for each different emotion. No further manual correction of the labels took place.

Due to recording and clipping problems, a few sentences from each emotion were discarded. Table 1 shows statistics on the post-processed corpus.[3]

6.3. Feature extraction

After processing the corpus, we resampled the audio at 16 kHz, and extracted 31 mel-frequency cepstral coefficients, the logarithmic fundamental frequency, and 25 band-aperiodicity coefficients with a frame shift of 5 ms using the STRAIGHT tool for MATLAB.

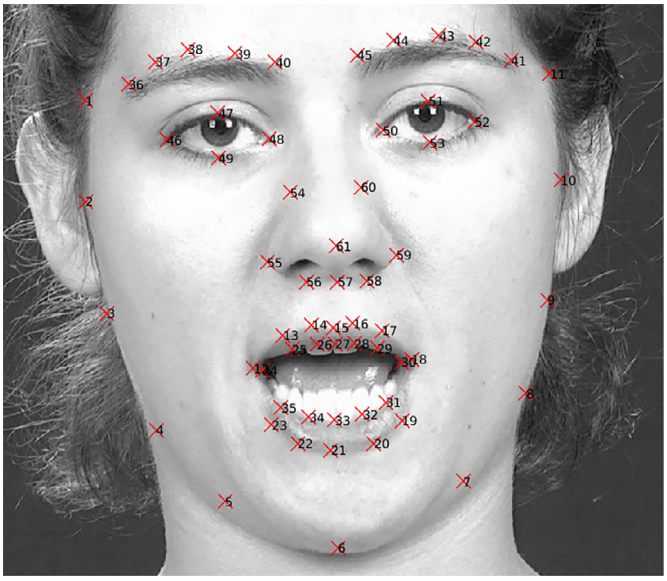

Fig. 8. Facial landmarks used for building the Active Appearance Model.

For the extraction of the eigenshape and eigentexture weights for each different frame in the database using the AAM modeling method described in Section 3, we used the following heuristic approach in order to minimize the fitting error of Eq. (4).

1. We hand labeled a total of 981 frames with 61 facial landmarks as shown in Fig. 8. From the 981 frames, 179 correspond to neutral expressions, 262 to angry expressions, 322 to happy expressions, and 218 to sad expressions.
2. For each different emotion in the corpus and its set of frames, we first use the face detection algorithm of Mathias et al. (2014) and then use multivariate regression to map from the detected rectangle to an initial shape that is go-

[3] The CVSP-EAV corpus is available at http://cvsp.cs.ntua.gr/research/eavtts/.

Table 2
Fitting results in terms of mean reconstruction error for each of the emotions in the CVSP-EAV Corpus.

	Neutral	Anger	Happiness	Sadness
Mean Rec. Error	0.0013	0.0013	0.0015	0.0013
# discarded	5	11	93	6

ing to be used to obtain an initial estimate for the first frame of each sentence:

$$\boldsymbol{S} = \boldsymbol{AR} + \boldsymbol{b} \tag{11}$$

where $\boldsymbol{S}$ is the shape corresponding to the detected rectangle $\boldsymbol{R}$, $\boldsymbol{A}$ is the regression matrix and $\boldsymbol{b}$ the intercepts. This results in four emotion-dependent regression models.
We also applied the same process to the whole corpus, in order to obtain an emotion-independent regression model. We found that generally the emotion-dependent regression models achieved a better shape initialization compared to the emotion-independent regression model.

3. We then proceeded to build an AAM using all of the annotated frames.
From the found eigenshapes and eigentextures, we keep in both cases the vectors that account for 95% of the variation, a total of 9 eigenshapes and 58 eigentextures.
4. For each different sentence in the corpus, we get the best initial estimate of the shape of the first frame in a sentence, by using either an emotion-dependent regression or an emotion-independent one. We choose by comparing the reconstruction errors after fitting with each different estimates. For each of the subsequent frames in the sentence, we use the shape found previously as an initial estimate.

In order to automatically exclude sentences where fitting presented artifacts, as much as possible, from the subsequent training of the systems, we discarded the sentences where the mean reconstruction error was above the threshold of 0.0018 and where more than 10 frames in the sentence had a reconstruction error of more than 0.0030. The threshold of 0.0018 was found heuristically to represent excellent fitting and reconstruction.

Table 2 contains the final mean reconstruction error resulting from fitting the Active Appearance Model for each different emotion and excluding the aforementioned sentences.

It is evident from both the mean reconstruction error as well as the number of sentences we had to discard, that the most difficult expression to fit is the expression of happiness which is expressed with strong variations in the human face. We would expect that the same would hold for the emotion of anger, however the number of discarded sentences for anger is not on the same level as with happiness.

The final visual features extracted for each sentence, were resampled at 200 fps in order to match the previously extracted acoustic features.

In the end, in order to have a fair comparison across all emotions in the experiments, we kept for training all sentences that were common across all emotions, a total of 774 sentences.

For annotating the frames we used the am-tools software[4] and for building and fitting the Active Appearance Model we use the AAM-tools toolkit[5] (Papandreou and Maragos, 2008).

7. Experimental results

7.1. Evaluation procedure

In order to assess the methods described in this paper we designed and developed a web-based questionnaire containing multiple types of questions and tests which will be described in the following sections. Each questionnaire[6] had a maximum of 102 random questions distributed to our different evaluations.

7.2. Evaluation of realism and expressiveness of the EAVTTS methods

Our evaluation of the EAVTTS methods described in this paper:

1. HMM-based EAVTTS (HMM)
2. DNN-based EAVTTS with joint modeling of acoustic and visual features (DNN-J)
3. DNN-based EAVTTS with separate modeling of acoustic and visual features (DNN-S)
4. Unit selection EAVTTS (US)

aims at comparing the methods both on the realism and expressiveness of the synthesized talking head. Furthermore, in order to gain more insight, we do not compare the methods only on the audio-visual realism, but also on the realism that is achieved by each different modality. "Realism" denotes the similarity of the talking head (or acoustic speech in case of the evaluation of the different modalities), to a human uttering the same sentence. This encapsulates both naturalness as well as intelligibility.

For each method, and for each of the four emotional training sets (neutral, anger, happiness, sadness) of the CVSP-EAV corpus we trained a subsystem (which we call from now on an emotion-independent AVTTS system). This means that, e.g., the full HMM-based EAVTTS system consists of 4 subsystems - one for each emotion. 48 test sentences, taken from the corpus, were generated from each subsystem. As a result $4 \times 48 = 192$ sentences were generated from each full EAVTTS method.

The HMM-based subsystems were built using the HTS toolkit (Zen et al., 2007a). Five state, context-dependent MSD-HSMMs with left-to-right topology were trained and tied using a decision-tree clustering technique, using a similar set of questions with Tokuda et al. (2002), but adapted for the Greek language. We use 29 different phonemes for the Greek language including silence.

Training of the DNN-based subsystems was implemented using the Merlin speech synthesis toolkit (Wu et al., 2016; Ronanki et al., 2016). The input vector to the neural networks was broken down to 494 linguistic features from the almost $\sim$1500 questions used for context clustering in the HMM-based systems, by exploiting the fact that non-binary linguistic features can be used as an input in the neural networks.

All networks (both networks that predict duration and networks that predict features) consisted of six hidden fully connected layers of 1024 neurons each. For training the networks with Back-propagation we use a batch size of 256 and a learning rate of 0.002. We train for a maximum of 25 epochs unless the error on the validation set (from the 774 sentences used to train each subsystem, we used 10 as a validation set) increases in more than 5 consecutive epochs after epoch 15. It is important to note at this stage that the architecture that employs separate modeling of acoustic and visual features, uses double the number of parameters than the architecture that employs joint modeling of acoustic and visual features.

[4] https://personalpages.manchester.ac.uk/staff/timothy.f.cootes/software/am_tools_doc/index.html.
[5] http://cvsp.cs.ntua.gr/software/AAMtools/.
[6] The questionnaire along with numerous videos of the talking head can be found at http://cvsp.cs.ntua.gr/research/eavtts/.

Table 3
Results (%) of subjective pairwise preference tests on audio-visual speech realism. Bold font indicates significant preference at $p < 0.01$ level.

DNN-S	DNN-J	HMM	US	N/P
25.0	22.22	–	–	52.78
51.11	–	15.56	–	33.33
75.56	–	–	18.89	5.55
–	**43.33**	22.22	–	34.44
–	**72.22**	–	22.78	5.0
–	–	**63.89**	27.78	8.33

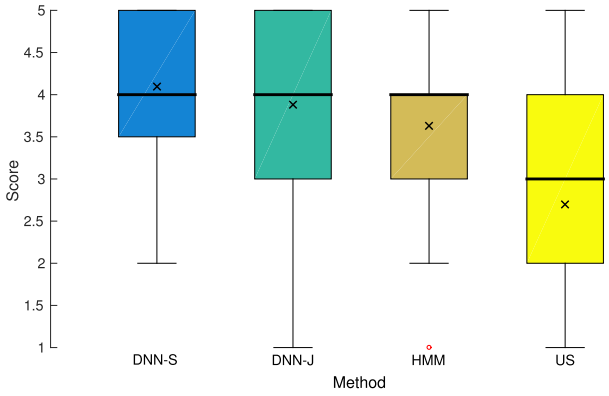

Fig. 9. Boxplot of the MOS test results on the audio-visual realism of the different EAVTTS methods. Bold line represents the median, *x* represents the mean, the boxes extend between the 1st and 3rd quantile, whiskers extend to the lowest and highest datum within 1.5 times the inter-quantile range of the 1st and 3rd quartile respectively, and outliers are represented with circles.

The unit selection subsystems were built by modifying an existing unit selection acoustic speech synthesis system, as described in Section 5.

7.2.1. Evaluation of audio-visual realism

To evaluate the realism of the talking head (both acoustic and visual) generated from each of the different methods, respondents of the web-based questionnaire were presented with pairs of videos depicting the video-realistic talking head uttering the same sentence and in the same emotion, generated by two of the previously described four methods, and were asked to choose the most realistic video in terms of both acoustic and visual streams (with a "no preference" option available as well). The sentences were chosen randomly from the 192 sentences that were generated from each system. We also made sure that all emotions appear in the same rate. The result is a total 6 pairwise preference tests (for all different combinations of the 4 methods), with 180 pairs evaluated for each method pair (45 pairs for each emotion). The results of the preference tests are presented in Table 3.

Our statistical analysis of preference tests employs a sign test (ignoring ties), with Holm–Bonferroni correction over all statistical tests of this section - 30 in total.

From the table we can see that both DNN architectures are preferred significantly at the $p < 0.01$ level over the HMM and US methods, while HMM is also preferred significantly over US at the $p < 0.01$ level. Among the two DNN architectures we see that the preference scores are very close and there is not a significant difference.

We generally observe a strong bias for the parametric approaches over the unit selection approach; a reasonable outcome considering that the size of each emotional training set is relatively low for unit selection synthesis combined with generation of unseen sentences.

Table 4
Significant differences between systems, from the MOS test results, on the audio-visual realism of the generated talking head, at levels $p < 0.05$ and $p < 0.01$. A blank cell denotes no significant difference.

	DNN-S	DNN-J	HMM	US
DNN-S	–		$p < 0.01$	$p < 0.01$
DNN-J	–	–	$p < 0.05$	$p < 0.01$
HMM	–	–	–	$p < 0.01$
US	–	–	–	–

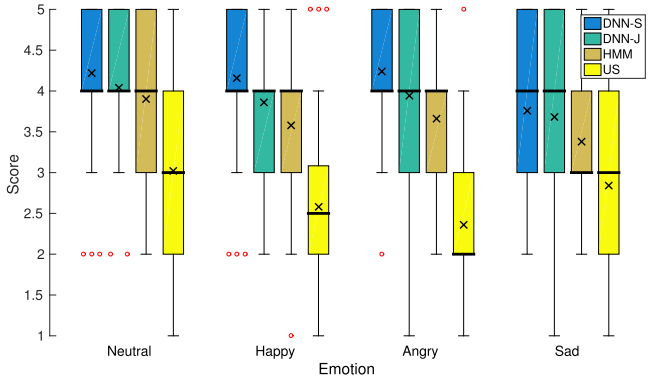

Fig. 10. Results of the MOS test broken down for each different emotion. Bold line represents the median, *x* represents the mean, the boxes extend between the 1st and 3rd quantile, whiskers extend to the lowest and highest datum within 1.5 times the inter-quantile range of the 1st and 3rd quartile respectively, and outliers are represented with circles.

A second evaluation of the audio-visual realism of the different methods was also performed, via a mean opinion score test (MOS). The respondents were presented with random videos of the talking head from each method and were asked to evaluate the realism of the talking head on a scale of 1 (poor realism) to 5 (perfect realism). Before the evaluation the respondents were also presented with samples from the original recordings and were instructed that they correspond to perfect realism. Each method was evaluated 200 times (50 for each emotion) and the results are shown in Fig. 9.

To check for significant differences between the systems we perform pairwise Mann–Whitney *U* tests (with the same Holm–Bonferroni correction as before) due to the fact that Likert-type scales are inherently ordinal scales (Clark et al., 2007). The results are shown in Table 4.

We can see that there is almost complete accordance of the results of the MOS test with the results obtained from the pairwise preference tests.

In Fig. 10 we also show the MOS test results for each different emotion.

7.2.2. Evaluation of visual realism

Similarly with the evaluation of audio-visual realism, we conducted 6 more pairwise preference tests in which respondents were presented with random pairs of muted videos and were

Table 5
Results (%) of subjective pairwise preference tests on visual speech realism. Bold font indicates significant preference at $p < 0.01$ level.

DNN-S	DNN-J	HMM	US	N/P
28.33	27.5	–	–	44.17
40.0	–	28.33	–	31.67
84.17	–	–	10.83	5.0
–	38.33	30.83	–	30.83
–	**85.0**	–	8.33	6.67
–	–	**76.67**	15.0	8.33

Table 6
Results (%) of subjective pairwise preference tests on acoustic speech realism. Bold font indicates significant preference at $p < 0.01$ level.

DNN-S	DNN-J	HMM	US	N/P
40.0	9.17	–	–	50.83
65.83	–	7.5	–	26.67
79.17	–	–	14.17	6.67
–	41.67	26.67	–	31.67
–	55.83	–	32.5	11.67
–	–	**64.17**	26.67	9.17

Table 7
Results (%) of subjective pairwise preference tests on audio-visual speech expressiveness. Bold font indicates significant preference at $p < 0.01$ level.

DNN-S	DNN-J	HMM	US	N/P
23.08	23.08	–	–	53.85
50.64	–	15.38	–	33.97
70.51	–	–	23.07	6.41
–	42.31	26.28	–	31.41
–	**66.67**	–	27.56	5.77
–	–	57.69	36.54	5.77

asked to pick the most realistic video (with a "no preference" option available). Each method pair was evaluated 120 times (30 pairs for each emotion), and the results are presented in Table 5.

From the table we can see that statistically significant differences occur only between parametric approaches versus the unit selection one. The DNN architectures seem again to be preferred over HMM, however the result is not statistically significant.

7.2.3. Evaluation of acoustic realism

For evaluating the acoustic speech generated, human evaluators were presented with random pairs of acoustic speech samples and were asked to pick the most realistic. Just like in the evaluation of visual realism, realism of acoustic speech was evaluated 120 times (30 pairs for each emotion) for each different method pair. The results are presented in Table 6.

We can see that all pairwise comparisons are significant at the $p < 0.01$ level, apart from the comparisons between DNN-J, HMM and DNN-J, US, where, although the DNN-J method is preferred, we did not observe statistical significance.

7.2.4. Evaluation of expressiveness

Expressiveness was evaluated in the same manner as audio-visual realism, were pairs of videos were presented and compared by human evaluators on their expressiveness. Videos of the neutral emotion were not included. The 6 pairwise preference tests on the evaluation of expressiveness were evaluated 156 (52 pairs for each emotion) times each, and we show the results in Table 7.

We see that the DNN-S architecture is significantly preferred over the HMM and US methods. The DNN-J architecture is significantly preferred over US, and although preferred over HMM, it is not significant in a statistical meaning. Between HMM and US, the former is preferred, though the result again is not statistically significant.

A correlation between realism and expressiveness is evident, since we can see that the results follow a resembling course with the evaluation of the audio-visual realism.

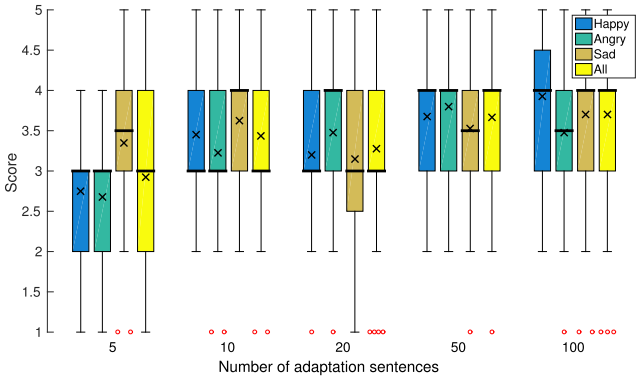

Fig. 11. Subjective evaluation of the level of expressiveness captured by an adapted HMM audio-visual speech synthesis system for each different emotion (and total), and for a variable number of sentences. Bold line represents the median, *x* represents the mean, the boxes extend between the 1st and 3rd quantile, whiskers extend to the lowest and highest datum within 1.5 times the inter-quantile range of the 1st and 3rd quartile respectively, and outliers are represented with circles.

7.3. Evaluation of HMM adaptation

To evaluate our second main focus, we adapted the HMM-based AVTTS subsystem trained on the neutral emotion of Part 7.2, to each of the other three emotions in the corpus, using the CSMAPLR adaptation described in Section 4.2, followed by a MAP adaptation. We also used a variable number of sentences for each of the above adaptations, namely 5, 10, 20, 50 and 100 sentences each time and for each of the different number of sentences, and emotions, we generated 8 unseen sentences from the test set.

In the questionnaire, each subject was presented with random videos for each different emotion (apart from neutral) and for each different number of adaptation sentences (a total of 15 videos), and were asked to evaluate the expressiveness of the talking head on an increasing scale of 1–5. We also included for each video, a video generated by the respective emotion individual HMM-based system built in Part 7.2 and advised the evaluators that this second video serves as a ground truth for the rating of 5, since we make the assumption that the adapted HMM system is capped, as far as expressiveness is concerned, by the corresponding emotion-independent HMM AVTTS system.

For each different emotion and number of adaptation sentences, 40 videos were evaluated (a total of 600 evaluations). Fig. 11 shows the results of this subjective evaluation, for each different emotion, and for each different number of sentences used for adaptation.

We observe that the median value over all emotions increases as the number of sentences used for adaptation increase. We also observe that the emotion of sadness achieves even for 5 adaptation sentences a large score/median, compared to the other two emotions. This could be explained by the fact that neutral speaking style possesses a similar speaking rate to the sad speaking style,

Table 8
Classification of emotions in the emotion individual HMM systems (% scores). Emotions not chosen by any respondent are not shown.

	Neutral	Happiness	Anger	Sadness	Fear	Pride	Pity	Other
Neutral	100.0	0	0	0	0	0	0	0
Happiness	0	80	0	0	0	13.33	0	6.67
Anger	6.67	0	73.33	6.67	0	0	6.67	6.67
Sadness	6.67	0	0	80.0	6.67	0	6.67	0

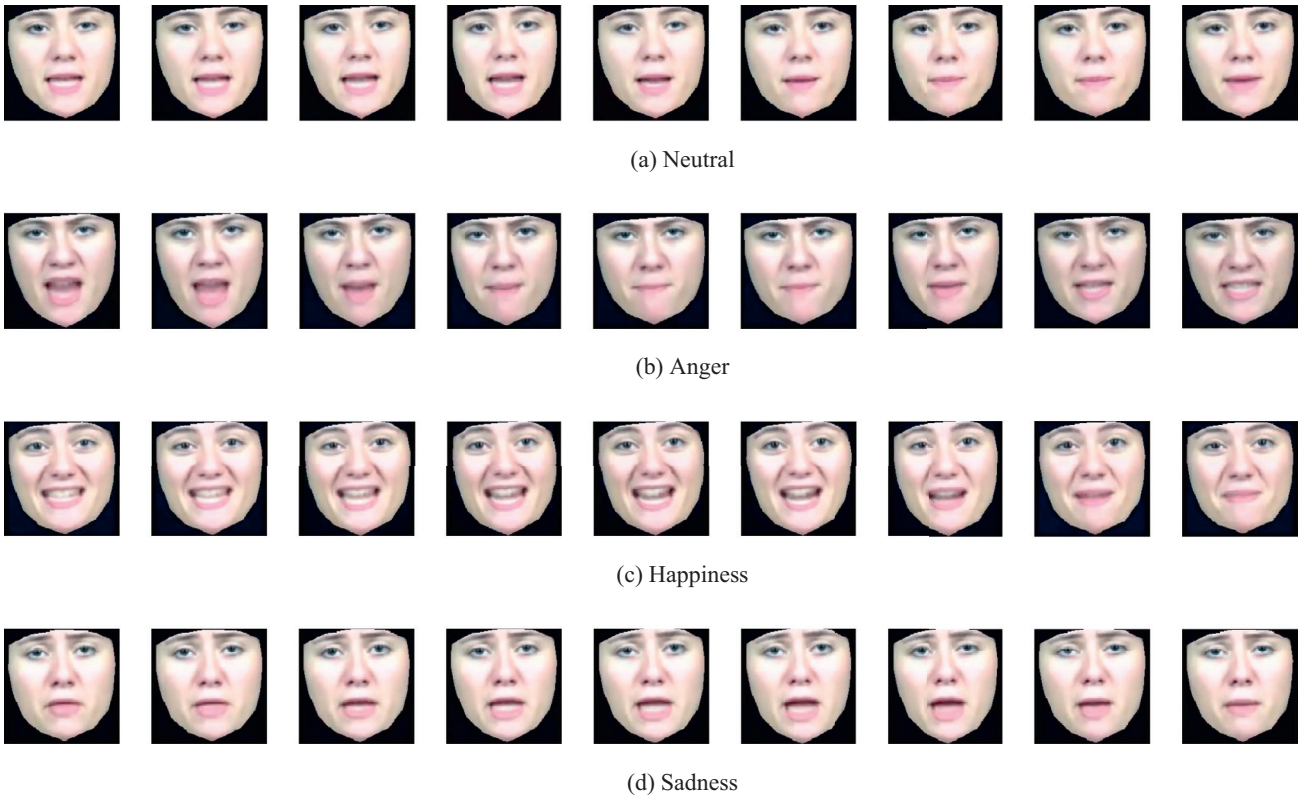

Fig. 12. Results of audio-visual synthesis (consecutive frames from the same sentence) from a neutral HMM set (a) and its adaptation to the three emotions of (b) anger, (c) happiness, and (d) sadness, using 50 adaptation sentences.

Table 9
Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting the neutral emotion, and the second one trained on an emotional training set depicting happiness (% scores, w_n: Neutral Weight, w_h: Happiness Weight). Emotions not chosen by any respondent are not shown.

	Emotions						
(w_n, w_h)	Neutral	Happiness	Sadness	Pride	Disgust	Pity	Other
(0.1, 0.9)	0	93.33	0	6.67	0	0	0
(0.3, 0.7)	13.33	80	6.67	0	0	0	0
(0.5, 0.5)	53.33	40.00	0	6.67	0	0	0
(0.7, 0.3)	66.67	00.00	0	6.67	6.67	6.67	13.33
(0.9, 0.1)	86.67	0	0	0	0	0	13.33

Table 10
Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting the neutral emotion, and the second one trained on an emotional training set depicting anger (% scores, w_n: Neutral Weight, w_a: Anger Weight). Emotions not chosen by any respondent are not shown.

	Emotions						
(w_n, w_a)	Neutral	Anger	Sadness	Pride	Disgust	Pity	Other
(0.1, 0.9)	13.33	66.67	0	6.67	6.67	0	6.67
(0.3, 0.7)	20.00	53.33	0	0	20	0	6.67
(0.5, 0.5)	46.67	33.33	0	6.67	13.33	0	0
(0.7, 0.3)	80.00	6.67	0	6.67	0	6.67	0
(0.9, 0.1)	86.67	0	6.67	6.67	0	0	0

as opposed to happiness and anger, where speaking rate is generally faster. It is important to note that we observe a high degree of agreement between the evaluators, since in almost all cases the range of the boxes is only 1 point on the MOS scale. Our general consensus is that HMM adaptation can be successfully employed for HMM-based EAVTTS.

In Fig. 12 we also show 10 consecutive frames from the same sentence, when adapting the neutral HMM set to one of the other three emotions using 50 sentences.

7.4. Evaluation of HMM interpolation

Finally, our final evaluation was on the application of HMM interpolation to the emotion individual HMM-based EAVTTS systems built in the first part of this section. As preparation, for each of the 6 different HMM set pairs arising when combining the 4 different emotions of our corpus, we generated 6 unseen sentences from the test set, using 5 sets of interpolation weights: (0.9, 0.1), (0.7, 0.3), (0.5, 0.5), (0.3, 0.7), (0.1, 0.9). We also generated the same sentences by each emotion individual EAVTTS system.

Next, respondents were presented with the generated videos of the talking head, and were asked to recognize the emotion depicted by choosing from a list containing 11 emotions (neutral, happiness, anger, sadness, fear, pride, surprise, disgust, pity, shame, envy), plus the “other” option.

As a first evaluation, and to show that our respondents indeed recognized the emotion corresponding to each emotion independent system built in Part 7.2, in Table 8 we show the results of emotion recognition for the emotion independent systems, where we can see all emotions achieve a high classification rate, with the lowest being anger with 73.33%. We note that each different inter-

Table 11
Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting the neutral emotion, and the second one trained on an emotional training set depicting sadness (% scores, w_n: Neutral Weight, w_s: Sadness Weight). Emotions not chosen by any respondent are not shown.

	Emotions								
(w_n, w_s)	Neutral	Sadness	Anger	Fear	Pride	Disgust	Pity	Envy	Other
(0.1, 0.9)	13.33	73.33	0	6.67	0	0	6.67	0	0
(0.3, 0.7)	20.00	73.33	0	0	0	0	6.67	0	0
(0.5, 0.5)	60.00	20.00	0	0	0	13.33	6.67	0	0
(0.7, 0.3)	73.33	6.67	0	0	6.67	6.67	0	0	6.67
(0.9, 0.1)	73.33	6.67	6.67	0	0	6.67	0	6.67	0

Table 12
Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting anger, and the second one trained on an emotional training set depicting happiness (% scores, w_a: Anger Weight, w_h: Happiness Weight). Emotions not chosen by any respondent are not shown.

	Emotions						
(w_a, w_h)	Neutral	Happiness	Anger	Pride	Disgust	Envy	Other
(0.1, 0.9)	0	86.67	0	13.33	0	0	0
(0.3, 0.7)	6.67	80.00	0	6.67	0	6.67	0
(0.5, 0.5)	13.33	60.00	0	13.33	6.67	0	6.67
(0.7, 0.3)	40.00	0	33.33	6.67	6.67	6.67	6.67
(0.9, 0.1)	0.0	0	86.67	0.0	6.67	0	6.67

Table 13
Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting anger, and the second one trained on an emotional training set depicting sadness (% scores, w_a: Anger Weight, w_s: Sadness Weight). Emotions not chosen by any respondent are not shown.

	Emotions								
(w_a, w_s)	Neutral	Anger	Sadness	Happiness	Fear	Disgust	Pity	Shame	Other
(0.1, 0.9)	6.67	0	80.00	6.67	0	6.67	0	0	0
(0.3, 0.7)	13.33	0	60	0	0	0	6.67	13.33	6.67
(0.5, 0.5)	33.33	33.33	13.33	0	6.67	0	6.67	0	6.67
(0.7, 0.3)	20.00	53.33	6.67	0	0	13.33	0	0	6.67
(0.9, 0.1)	6.67	66.67	0.0	0	0	20	0	0	6.67

Table 14
Emotion classification rate when interpolating two HMM sets; the first one trained on an emotional training set depicting sadness, and the second one trained on an emotional training set depicting happiness (% scores, w_s: Sadness Weight, w_h: Happiness Weight). Emotions not chosen by any respondent are not shown.

	Emotions										
(w_s, w_h)	Neutral	Happiness	Sadness	Anger	Fear	Pride	Disgust	Pity	Shame	Envy	Other
(0.1, 0.9)	6.67	73.33	6.67	0	0	6.67	6.67	0	0	0	0
(0.3, 0.7)	6.67	26.67	26.67	0	0	0	6.67	0	0	6.67	20.0
(0.5, 0.5)	20.00	6.67	33.33	0	6.67	6.67	0	13.33	6.67	0	6.67
(0.7, 0.3)	13.33	0	60.00	6.67	6.67	0	0	6.67	0	0	6.67
(0.9, 0.1)	0.00	0	100.0	0	0	0	0	0	0	0	0

polation pair and emotion-independent system was evaluated 15 times.

Subsequently, we present 6 tables that show the emotion classification rate for each of the emotion pairs and interpolation pairs, in Tables 9–14 (in the tables we only include the emotions which were picked - that is classification rate was above zero at least in one row).

A study of Tables 9–11, reveals that we can indeed achieve different levels of the emotions by their interpolation with the neutral emotion since the emotion recognition results fluctuate mainly between the neutral emotion and the emotion under consideration in each figure. Abrupt changes in the classification scores suggest that we need to have an even smaller interpolation step, in order to control the resulting intensity level.

In Tables 12–14 we can see the same trend. It is interesting to note, that the “neutral” emotion was also chosen many times. This result might suggest that the level of expressiveness at a weight of 0.5 is not strong enough, and when interpolated with another emotion at the same level the confusion causes the viewers to select the neutral stream. Several other options were also selected. We can see that for specific pairs, audiovisual speech with intermediate speaking style is generated (Anger-Sadness with respective weights (0.5, 0.5) and Sadness-Happiness with respective weights (0.3, 0.7)). We believe that a further study with more refined steps between the weights is imperative.

In Fig. 13 we also show 10 consecutive frames from interpolating the HMM sets trained with the emotions of happiness and anger, for the weights we previously stated.

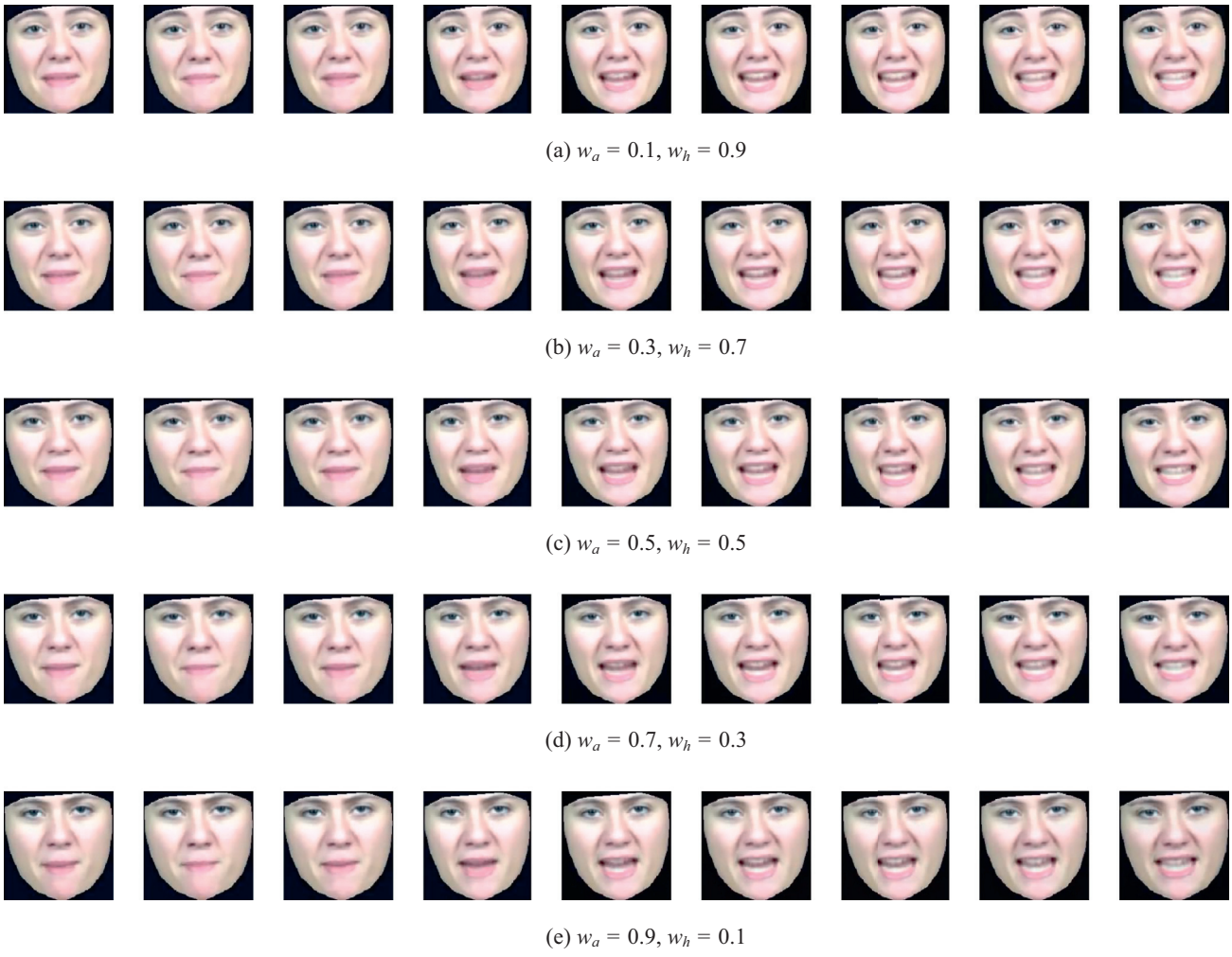

Fig. 13. Results of audio-visual synthesis (consecutive frames from the same sentence) from interpolating HMM sets trained on anger and happiness (w_a: anger weight, w_h: happiness weight).

8. Conclusion

In this paper, we performed a much-needed in-depth study on video-realistic expressive audio-visual speech synthesis, in order to improve this area through facing the challenges it poses. Towards that goal, we proposed two different architectures for DNN-based expressive audio-visual speech synthesis and did a direct comparison with HMM-based and concatenative unit selection expressive audio-visual speech synthesis systems on the realism of the produced talking head, and on the emotional strength that is captured by each system when it is trained on an emotional corpus.

Our results show that both DNN-based architectures significantly outperform the other two methods in terms of the audio-visual realism of the synthesized talking head, while the DNN-based architecture that uses separate modeling of acoustic and visual features architecture (DNN-S) significantly outperforms the HMM and US methods in terms of expressiveness as well. In addition, DNN-S also achieved significantly better results over all other architectures when considering acoustic speech only.

The results of the unit selection system were much worse in comparison with parametric approaches, which is to be expected when considering not only the fact that our corpus is fairly small for US synthesis, but also that the number of needed units increases when considering expressive speech.

In addition, we adopted CSMAPLR adaptation in order to adapt an HMM system to a target emotion using a small number of adaptation sentences, and showed that adaptation can successfully be applied to EAVTTS. We also showed the fact that HMM interpolation can be employed in order to achieve different levels of intensity for the emotions of our corpus, but also expressions and speech with intermediate speaking styles. Our last contribution is a medium sized audio-visual speech corpus for the Greek language, featuring three emotions: anger, happiness, and sadness, plus the neutral reading style.

We believe that our study opens multiple directions for future work. It would be interesting to compare RNN architectures and their variations with our methods, always on an expressive speech ground. Furthermore, since DNN-based architectures outperform HMM-based architectures, it is imperative to research adaptation and interpolation of DNN-based EAVTTS systems in order to tackle the challenges we stated that arise when considering expressive audio-visual speech.

Acknowledgments

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

The authors wish to thank Dimitra Tarousi for her participation in the recordings of the CVSP-EAV corpus.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.specom.2017.08.011.

References

Ambady, N., Rosenthal, R., 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychol. Bull. 111, 256–274.

Anderson, R., Stenger, B., Wan, V., Cipolla, R., 2013. Expressive visual text-to-speech using active appearance models. In: Proc. CVPR., pp. 3382–3389.

Bailly, G., Bérar, M., Elisei, F., Odisio, M., 2003. Audiovisual speech synthesis. Intl. J. Speech Technol. 6, 331–346.

Bates, J., 1994. The role of emotion in believable agents. Commun. ACM 37, 122–125.

Bechara, A., 2004. The role of emotion in decision-making: evidence from neurological patients with orbitofrontal damage. Brain Cogn. 55, 30–40.

Beskow, J., 1996. Talking heads – communication, articulation and animation. In: Proc. Fonetik. Nasslingen, Sweden.

Black, A.W., 2003. Unit selection and emotional speech. In: Proc. Interspeech, pp. 1649–1652.

Cao, Y., Tien, W.C., Faloutsos, P., Pighin, F., 2005. Expressive speech-driven facial animation. ACM Trans. Graph. 24, 1283–1302.

Chalamandaris, A., Tsiakoulis, P., Karabetsos, S., Raptis, S., LTD, I., 2013. The ILSP/INNOETICS text-to-speech system for the Blizzard Challenge 2013. In: Proc. Blizzard Challenge.

Clark, R.A., Podsiadlo, M., Fraser, M., Mayo, C., King, S., 2007. Statistical analysis of the Blizzard Challenge 2007 listening test results. In: Proc. ISCA SSW6.

Cootes, T.F., Edwards, G.J., Taylor, C.J., 2001. Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. 23, 681–685.

Cosatto, E., Potamianos, G., Graf, H.P., 2000. Audio-visual unit selection for the synthesis of photo-realistic talking-heads. In: Proc. ICME, vol. 2, pp. 619–622.

Darwin, C., 1871. The Expression of the Emotions in Man and Animals.

Deng, Z., Neumann, U., Lewis, J.P., Kim, T.-Y., Bulut, M., Narayanan, S., 2006. Expressive facial animation synthesis by learning speech coarticulation and expression spaces. IEEE Trans. Vis. Comput. Graph. 12, 1523–1534.

Digalakis, V., Oikonomidis, D., Pratsolis, D., Tsourakis, N., Vosnidis, C., Chatzichrisafis, N., Diakoloukas, V., 2003. Large vocabulary continuous speech recognition in Greek: corpus and an automatic dictation system. In: Proc. Interspeeech, pp. 1565–1568.

Digalakis, V.V., Neumeyer, L.G., 1996. Speaker adaptation using combined transformation and Bayesian methods. IEEE Trans. Speech Audio Process 4, 294–300.

Ekman, P., 1984. Expression and the nature of emotion. In: Approaches to Emotion, vol. 3, pp. 19–344.

Ekman, P., Freisen, W.V., Ancoli, S., 1980. Facial signs of emotional experience. J. Pers. Soc. Psychol. 39 (6), 1125.

Ezzat, T., Geiger, G., Poggio, T., 2002. Trainable videorealistic speech animation. In: Proc. ACM SIGGRAPH, pp. 388–398.

Fan, B., Xie, L., Yang, S., Wang, L., Soong, F.K., 2015. A deep bidirectional LSTM approach for video-realistic talking head. Multimed. Tools Appl. 75, 5287–5309.

Gales, M.J., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. 12 (2), 75–98.

Hatfield, E., Cacioppo, J.T., Rapson, R.L., 1993. Emotional contagion. Curr. Dir. Psychol. Sci. 2, 96–100.

Heiga, Z., Tokuda, K., Masuko, T., Kobayasih, T., Kitamura, T., 2007. A hidden semi-Markov model-based speech synthesis system. IEICE Trans. Inf. Syst. 90, 825–834.

Hess, U., Blairy, S., Kleck, R.E., 1997. The intensity of emotional facial expressions and decoding accuracy. J. Nonverbal Behav. 21 (4), 241–257.

Huang, F.J., Cosatto, E., Graf, H.P., 2002. Triphone based unit selection for concatenative visual speech synthesis. In: Proc. ICASSP, vol. 2, pp. 2037–2040.

Johnson, W.L., Rickel, J.W., Lester, J.C., 2000. Animated pedagogical agents: face-to-face interaction in interactive learning environments. J. Artif. Intell. Educ. 11, 47–78.

Jones, M.J., Poggio, T., 1998. Multidimensional morphable models. In: Proc. ICCV, pp. 683–688.

Katsamanis, A., Black, M., Georgiou, P.G., Goldstein, L., Narayanan, S., 2011. SailAlign: Robust long speech-text alignment. In: Proc. VLSP.

Kawahara, H., Estill, J., Fujimura, O., 2001. 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.

Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A., 1999. 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. 197–207.

Keltner, D., Haidt, J., 1999. Social functions of emotions at four levels of analysis. Cogn. Emot. 13, 505–521.

Le Goff, B., Benoît, C., 1996. A text-to-audiovisual-speech synthesizer for French. In: Proc. ICSLP, vol. 4, pp. 2163–2166.

Leggetter, C., Woodland, P., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput. Speech Lang. 9 (2), 171–185.

Li, X., Wu, Z., Meng, H., Jia, J., Lou, X., Cai, L., 2016. Expressive speech driven talking avatar synthesis with DBLSTM using limited amount of emotional bimodal data. In: Proc. Interspeech, pp. 1477–1481.

Ling, Z.-H., Kang, S.-Y., Zen, H., Senior, A., Schuster, M., Qian, X.-J., Meng, H.M., Deng, L., 2015. Deep learning for acoustic modeling in parametric speech generation: a systematic review of existing techniques and future trends. IEEE Signal Process. Mag. 32, 35–52.

Liu, K., Ostermann, J., 2011. Realistic facial expression synthesis for an image-based talking head. In: Proc. ICME, pp. 1–6.

Lorenzo-Trueba, J., Barra-Chicote, R., San-Segundo, R., Ferreiros, J., Yamagishi, J., Montero, J.M., 2015. Emotion transplantation through adaptation in HMM-based speech synthesis. Comput. Speech Lang. 34, 292–307.

Masuko, T., Tokuda, K., Kobayashi, T., Imai, S., 1997. Voice characteristics conversion for HMM-based speech synthesis system. In: Proc. ICASSP, vol. 3, pp. 1611–1614.

Mathias, M., Benenson, R., Pedersoli, M., Van Gool, L., 2014. Face detection without bells and whistles. In: Proc. ECCV, pp. 720–735.

Matthews, I., Baker, S., 2004. Active appearance models revisited. Int. J. Comput. Vis. 60, 135–164.

Mattheyses, W., Latacz, L., Verhelst, W., 2011. Auditory and photo-realistic audiovisual speech synthesis for Dutch.. In: Proc. AVSP, pp. 55–60.

Mattheyses, W., Latacz, L., Verhelst, W., Sahli, H., 2008. Multimodal unit selection for 2d audiovisual text-to-speech synthesis. In: Proc. MLMI, pp. 125–136.

Mattheyses, W., Verhelst, W., 2015. Audiovisual speech synthesis: an overview of the state-of-the-art. Speech Commun. 66, 182–217.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748.

Melenchón, J., Martínez, E., De La Torre, F., Montero, J.A., 2009. Emphatic visual speech synthesis. IEEE Trans. Audio Speech Lang. Process 17, 459–468.

Mori, M., MacDorman, K.F., Kageki, N., 2012. The uncanny valley [from the field]. IEEE Robot. Autom. Mag. 19, 98–100.

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

Ouni, S., Cohen, M.M., Ishak, H., Massaro, D.W., 2007. Visual contribution to speech perception: measuring the intelligibility of animated talking heads. EURASIP J. Audio Speech Music Process 2007, 3.

Pandzic, I.S., Forchheimer, R., 2003. MPEG-4 Facial Animation: the Standard, Implementation and Applications.

Papandreou, G., Maragos, P., 2008. Adaptive and constrained algorithms for inverse compositional active appearance model fitting. In: Proc. CVPR, pp. 1–8.

Pelachaud, C., Badler, N.I., Steedman, M., 1996. Generating facial expressions for speech. Cogn. Sci. 20 (1), 1–46.

Pérez, P., Gangnet, M., Blake, A., 2003. Poisson image editing. ACM Trans. Graph. 22, 313–318.

Plutchik, R., 1980. Emotion: A Psychoevolutionary Synthesis.

Plutchik, R., 2001. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am. Sci. 89, 344–350.

Plutchik, R., Kellerman, H., 1980. Emotion, Theory, Research, and Experience: Theory, Research and Experience.

Qian, Y., Fan, Y., Hu, W., Soong, F.K., 2014. On the training aspects of deep neural network (DNN) for parametric TTS synthesis. In: Proc. ICASSP, pp. 3829–3833.

Raptis, S., Tsiakoulis, P., Chalamandaris, A., Karabetsos, S., 2016. Expressive speech synthesis for storytelling: the INNOETICS' entry to the blizzard challenge 2016. In: Proc. Blizzard Challenge.

Richard, J.D., Klaus, R.S., Goldsmith, H.H., 2002. Handbook of Affective Sciences.

Ronanki, S., Wu, Z., Watts, O., King, S., 2016. A demonstration of the Merlin open source neural network speech synthesis system. In: Proc. ISCA SSW9.

Russell, S., Norvig, P., 1995. Artificial Intelligence: A Modern Approach.

Sako, S., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2000. HMM-based text–to-audio-visual speech synthesis. In: Proc. ICLSP, pp. 25–28.

Salvi, G., Beskow, J., Al Moubayed, S., Granström, B., 2009. SynFace - speech-driven facial animation for virtual speech-reading support. EURASIP J. Audio Speech Music Process. 2009, 191940.

Schabus, D., Pucher, M., Hofer, G., 2014. Joint audiovisual hidden semi-Markov model-based speech synthesis. IEEE J. Sel. Topics Signal Process. 8, 336–347.

Schröder, M., 2009. Expressive speech synthesis: past, present, and possible futures. In: Affective information processing. Springer, pp. 111–126.

Schwarz, N., 2000. Emotion, cognition, and decision making. Cogn. Emot. 14, 433–440.

Seyama, J., Nagayama, R.S., 2007. The uncanny valley: effect of realism on the impression of artificial human faces. Presence 16, 337–351.

Shaw, F., Theobald, B.-J., 2016. Expressive modulation of neutral visual speech. IEEE Multimedia 23, 68–78.

Shinoda, K., Lee, C.-H., 2001. A structural Bayes approach to speaker adaptation. IEEE Trans. Speech Audio Process 9, 276–287.

Shinoda, K., Watanabe, T., 1997. Acoustic modelling based on the mdl principle for speech recognition. In: Proc. Eurospeech, pp. 99–102.

Sifakis, E., Selle, A., Robinson-Mosher, A., Fedkiw, R., 2006. Simulating speech with a physics-based facial muscle model. In: Proc. ACM SIGGRAPH, pp. 261–270.

Skipper, J.I., van Wassenhove, V., Nusbaum, H.C., Small, S.L., 2007. Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. Cereb. Cortex 17, 2387–2399.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212–215.

Tachibana, M., Yamagishi, J., Masuko, T., Kobayashi, T., 2005. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. IEICE Trans. Inf. Syst. 88, 2484–2491.

Tamura, M., Kondo, S., Masuko, T., Kobayashi, T., 1999. Text-to-audiovisual speech synthesis based on parameter generation from HMM. In: Proc. Eurospeech, p. 959962.
Tamura, M., Masuko, T., Tokuda, K., Kobayashi, T., 2001. Adaptation of pitch and spectrum for HMM-based speech synthesis using mllr. In: Proc. ICASSP, 2, pp. 805–808.
Tokuda, K., Toda, T., Yamagishi, J., 2013. Speech synthesis based on hidden Markov models. Proc. IEEE 101, 1234–1252.
Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In: Proc. ICASSP, vol. 3, pp. 1315–1318.
Tokuda, K., Zen, H., Black, A.W., 2002. An HMM-based speech synthesis system applied to English. In: Proc. IEEE SSW, pp. 227–230.
Turk, M.A., Pentland, A.P., 1991. Face recognition using eigenfaces. In: Proc. CVPR, pp. 586–591.
Wan, V., Anderson, R., Blokland, A., Braunschweiler, N., Chen, L., Kolluru, B., Latorre, J., Maia, R., Stenger, B., Yanagisawa, K., Stylianou, Y., Akamine, M., Gales, M., Cipolla, R., 2013. Photo-realistic expressive text to talking head synthesis. In: Proc. Interspeech, pp. 2667–2669.
Wang, L., Qian, X., Han, W., Soong, F.K., 2010. Photo-real lips synthesis with trajectory-guided sample selection. In: Proc. ISCA SSW7, pp. 217–222.
Watts, O., Henter, G.E., Merritt, T., Wu, Z., King, S., 2016. From HMMs to DNNs: where do the improvements come from? In: Proc. ICASSP, vol. 41, pp. 5505–5509.
Williams, D., Hinton, G., 1986. Learning representations by back-propagating errors. Nature 323, 533–536.
Wu, Z., Watts, O., King, S., 2016. Merlin: An open source neural network speech synthesis system. In: Proc. ISCA SSW9.
Wu, Z., Zhang, S., Cai, L., Meng, H.M., 2006. Real-time synthesis of chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar.. In: Proc. Interspeech.
Xie, L., Liu, Z.-Q., 2007. A coupled HMM approach to video-realistic speech animation. Pattern Recognit. 40, 2325–2340.
Xie, L., Sun, N., Fan, B., 2014. A statistical parametric approach to video-realistic text-driven talking avatar. Multimed. Tools Appl. 73, 377–396.
Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J., 2009. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio, Speech, Lang. Process. 17, 66–83.
Yamagishi, J., Masuko, T., Kobayashi, T., 2004. HHMM-based expressive speech synthesis - towards TTS with arbitrary speaking styles and emotions. In: Proc. of Special Workshop in Maui.
Yamagishi, J., Zen, H., Toda, T., Tokuda, K., 2007. Speaker-independent HMM-based speech synthesis system: HTS-2007 system for the Blizzard Challenge 2007. In: Proc. Blizzard Challenge.
Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In: Proc. Eurospeech, pp. 2347–2350.
Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speaker interpolation for HMM-based speech synthesis system. Acoust. Sci. Technol. 21, 199–206.
Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2005. Incorporating a mixed excitation model and postfilter into HMM-based text-to-speech synthesis. Syst. Comput. Jpn 36, 43–50.
Zen, H., 2015. Acoustic modeling in statistical parametric speech synthesis–from HMM to LSTM-RNN. In: Proc. MLSLP. Invited Paper.
Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., Tokuda, K., 2007. The HMM-based speech synthesis system (HTS) version 2.0. In: Proc. ISCA SSW6, pp. 294–299.
Zen, H., Senior, A., Schuster, M., 2013. Statistical parametric speech synthesis using deep neural networks. In: Proc. ICASSP, pp. 7962–7966.
Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Commun. 51, 1039–1064.
Zen, H., Tokuda, K., Masuko, T., Kobayasih, T., Kitamura, T., 2007. A hidden semi–Markov model-based speech synthesis system. IEICE Trans. Inf. Syst. E90-D, 825–834.