Expressive Visual Text-To-Speech Using Active Appearance Models

Robert Anderson[1] Björn Stenger[2] Vincent Wan[2] Roberto Cipolla[1]
[1] *Department of Engineering, University of Cambridge, Cambridge, UK*
[2] *Toshiba Research Europe, Cambridge, UK*

Abstract

This paper presents a complete system for expressive visual text-to-speech (VTTS), which is capable of producing expressive output, in the form of a 'talking head', given an input text and a set of continuous expression weights. The face is modeled using an active appearance model (AAM), and several extensions are proposed which make it more applicable to the task of VTTS. The model allows for normalization with respect to both pose and blink state which significantly reduces artifacts in the resulting synthesized sequences. We demonstrate quantitative improvements in terms of reconstruction error over a million frames, as well as in large-scale user studies, comparing the output of different systems.

1. Introduction

This paper presents a system for expressive visual text-to-speech (VTTS) that generates near-videorealistic output. Given an input text, a visual text-to-speech system generates a video of a synthetic character uttering the text. *Expressive* VTTS allows the text to be annotated with emotion labels which modulate the expression of the generated output. Creating and animating talking face models with a high degree of realism has been a long-standing goal, as it has significant potential for digital content creation and enabling new types of user interfaces [16, 19, 27]. It is becoming increasingly clear that in order to achieve this aim, one needs to draw on methods from different areas, including computer graphics, speech processing, and computer vision. While systems exist that produce high quality animations for neutral speech [6, 16, 25], adding controllable, realistic facial expressions is still challenging [1, 5]. Currently the most realistic data-driven VTTS systems are based on unit selection, splitting up the video into short sections and subsequently concatenating and blending these sections at the synthesis stage, *e.g.* [16, 25]. Due to the high degree of variation in appearance during expressive speech, the number of units required to allow realistic animation becomes excessive.

In our approach we draw on recent progress from the area of audio-only text-to-speech (TTS), which also has to deal with *coarticulation*, whereby phonemes are affected by other nearby phonemes. The most successful approach to solving this task currently is to model tri- or quinphones using hidden Markov models (HMM) with three or five emitting states, respectively [29]. Concatenating the HMMs and sampling from them produces a set of parameters which can then be synthesized into a speech signal. In order to extend this approach to visual TTS, a parametric face model is required. In this paper we propose using the established active appearance model (AAM) to model face shape and appearance [7]. While AAMs have been used in VTTS systems for neutral speech in the past [10, 23], there are a number of difficulties when applying standard AAMs to the task of expressive face modeling. The most significant problem is that AAMs capture a mixture of expression, mouth shape and head pose within each mode, making it impossible to model these effects independently. Due to the large variation of pose and expression in expressive VTTS this leads to artifacts in synthesis as spurious correlations are learned. AAMs are also inherently poor at modeling very localized actions such as blinking, without introducing artifacts elsewhere in the model when used for synthesis. In this paper we propose a number of extensions that allow AAMs to be used for synthesis tasks with a higher degree of realism. In summary, the contributions of this paper are:

1. a complete visual text-to-speech system allowing synthesis with a continuous range of emotions, introduced in section 4,

2. extensions to the standard AAM that allow the separation of modes for global and local shape and appearance deformations, detailed in section 3, and

3. large-scale, crowd-sourced user studies, allowing a direct comparison of the proposed system with the state of the art, see section 5. The experiments demonstrate a clear improvement in synthesis quality in expressive VTTS.

2. Prior Work

This section gives an overview of recent approaches to visual text-to-speech, grouping them based on their generative model.

Physics based methods model face movement based on simulating the effects of muscle interaction, thereby allowing anatomically plausible animation [1, 20]. However building accurate models requires significant effort and results are currently not videorealistic.

Unit selection methods allow videorealistic synthesis as they concatenate examples actually seen in a training set [16, 25]. The type of unit can be a single frame for each phoneme [12] or a sequence of frames [5] which are blended with their temporal neighbors. The advantage of longer units is that they better model coarticulation, however more units are required in this case to handle all phoneme combinations. The main drawback of unit selection approaches is their lack of flexibility, as they cannot easily be extended to handle new expressions without greatly increasing the number of units.

Statistical modeling approaches use a training set to build models of the speech generation process. HMMs are currently the most popular approach [4, 8]. Statistical models are able to generate high quality results which are sometimes over-smoothed compared to unit selection approaches. The main advantages of these methods are the flexibility they provide in dealing with coarticulation and their ability to handle expression variation in a principled manner.

2.1. Face models for VTTS

A number of different face models have been proposed for videorealistic VTTS systems.

Image based models use complete or partial images taken directly from a training set, concatenating them using warping or blending techniques. The resulting appearance is realistic, but this technique limits the synthesis method to unit selection [16, 26].

Data-driven 3D models use captured 3D data to generate controllable 3D models. Their main advantages are their invariance to 3D pose changes and their ability to render with an arbitrary pose and lighting at synthesis time. Currently a limiting factor is the complexity of the capture and registration process. While computer vision techniques continue to drive progress in this area [3, 17], until now only relatively small training sets have been acquired, insufficient in size to generate realistic expressive models [5, 24]. Good results have been achieved animating 3D models that do not attempt to appear videorealistic, this avoids the uncanny valley and produces visually appealing synthesis such as that in [21].

Data-driven 2D models can be created from video data, thereby simplifying the capture process of large training corpora. The most common 2D models used are AAMs [10, 23] and Multidimensional Morphable Models (MMMs) [6]. Both of these models are linear in both shape and appearance, but while AAMs represent shape using the position of mesh vertices, MMMs use flow fields to represent 2D deformation.

2.2. Active appearance models

In this paper we use AAMs as they produce good results for neutral speech while the low-dimensional parametric representation enables their combination with standard TTS methods. There have been many modifications to the standard AAM designed to target specific applications, see [13] for an overview. The specific requirements for our system are that the model must be able to track robustly and quickly over a very large corpus of expressive training data and that it must be possible to synthesize videorealistic renderings from statistical models of its parameters. There has been extensive work on tracking expressive data, for example the work of De la Torre and Black [9] in which several independent AAMs representing different regions of the face are created by hand are linked together by a shared affine warp. Modifications for convincing synthesis from AAMs on the other hand are much less well explored. When AAMs have been used for VTTS in the past, small head pose variations have been removed by subtracting the mean AAM parameters for each sentence from all frames within that sentence [10] however this approach works for small rotations only and leads to a loss of expressiveness. Bilinear AAMs that factor out pose from other motion have been proposed, but the amount of training data required for a VTTS system makes their use prohibitive in our application [14]. The most similar approach to dealing with pose to the method that we propose is that of Edwards *et al*. [11] in which canonical discriminant analysis is used to find semantically meaningful modes and a least squares approach is used to remove the contributions of these modes from training samples. However this approach is not well suited to modeling local deformations such as blinking and the least squares approach to removing the learned modes from training samples can give disproportionate weighting to the appearance component.

3. Extending AAMs for Expressive Faces

This section first briefly introduces the standard AAM with its notation and then details the proposed extensions to improve its performance in the expressive VTTS setting. As a baseline we use the AAM proposed by Cootes *et al*. [7] in which a single set of parameters controls both shape and appearance. Throughout this paper we assume that the number of shape and appearance modes is equal but the techniques are equally applicable if this is not the case; modes with zero magnitude can be inserted to en-

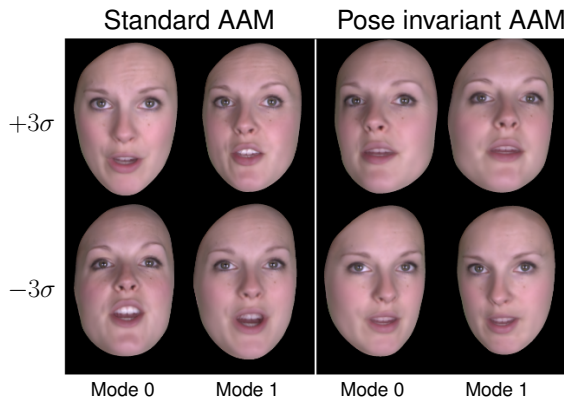


Figure 1: Pose invariant AAM modes. *The first two modes of a standard AAM* **(left)** *encode a mixture of pose, mouth shape and expression variation.* **(right)** *The first two modes of a pose invariant AAM encode only rotation, allowing head pose to be decoupled from expression and mouth shape.*

sure that the number of modes is equal. An AAM is defined on a mesh of V vertices. The shape of the model, $\mathbf{s} = (x_1, y_1, x_2, y_2, ...x_V, y_V)^T$, defines the 2D position (x_i, y_i) of each mesh vertex and is a linear model given by

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{M} c_i \mathbf{s}_i, \tag{1}$$

where $\mathbf{s_0}$ is the mean shape of the model, $\mathbf{s}_i$ is the ith mode of M linear shape modes and c_i is its corresponding parameter. We include color values in the appearance of the model, which is given by $\mathbf{a} = (r_1, g_1, b_1, r_2, g_2, b_2, ...r_P, g_P, b_P)^T$, where (r_i, g_i, b_i) is the RGB representation of the ith of the P pixels which project into the mean shape $\mathbf{s}_0$. Analogous to the shape model, the appearance is given by

$$\mathbf{a} = \mathbf{a}_0 + \sum_{i=1}^{M} c_i \mathbf{a}_i, \tag{2}$$

where $\mathbf{a}_0$ is the mean appearance vector of the model, and $\mathbf{a}_i$ is the ith appearance mode. Since we use a combined appearance model the weights c_i in equations 1 and 2 are the same and control both shape and appearance.

3.1. Pose invariant AAM modes

The global nature of AAMs leads to some of the modes handling variation which is due to both 3D pose change as well as local deformation, see figure 1 left. Here we propose a method for finding AAM modes that correspond purely to head rotation or to other physically meaningful motions. More formally, we would like to express a face shape $\mathbf{s}$ as a combination of pose components and deformation components:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{K} c_i \mathbf{s}_i^{\text{pose}} + \sum_{i=K+1}^{M} c_i \mathbf{s}_i^{\text{deform}} . \tag{3}$$

We would also like to obtain the equivalent expression for the appearance. The coupling of shape and appearance in AAMs makes this a difficult problem. We first find the shape components that model pose $\{\mathbf{s}_i^{\text{pose}}\}_{i=1}^K$, by recording a short training sequence of head rotation with a fixed neutral expression and applying PCA to the observed mean normalized shapes $\hat{\mathbf{s}} = \mathbf{s} - \mathbf{s}_0$. We then project $\hat{\mathbf{s}}$ into the pose variation space spanned by $\{\mathbf{s}_i^{\text{pose}}\}_{i=1}^K$ to estimate the weights $\{c_i\}_{i=1}^K$ in (3):

$$c_i = \frac{\hat{\mathbf{s}}^{\text{T}} \mathbf{s}_i^{\text{pose}}}{||\mathbf{s}_i^{\text{pose}}||^2} . \tag{4}$$

Having found these weights we remove the pose component from each training shape to obtain a pose normalized training shape $\mathbf{s}^*$:

$$\mathbf{s}^* = \hat{\mathbf{s}} - \sum_{i=1}^{K} c_i \mathbf{s}_i^{\text{pose}} . \tag{5}$$

If shape and appearance were indeed independent then we could find the deformation components by principal component analysis (PCA) of a training set of shape samples normalized as in (5), ensuring that only modes orthogonal to the pose modes are found, in the same way as [11]. However, there is no guarantee that the weights calculated using (4) are the same for the shape and appearance modes, which means that we may not be able to reconstruct the training examples using the model. This can be problematic, for example if the original AAM tracking method proposed in [7] or the method introduced in section 3.4 are to be used, as these require the AAM descriptors for each training sample. To overcome this problem we compute the mean of each $\{c_i\}_{i=1}^K$ of the appearance and shape weights:

$$c_i = \frac{1}{2}\left(\frac{\hat{\mathbf{s}}^{\text{T}} \mathbf{s}_i^{\text{pose}}}{||\mathbf{s}_i^{\text{pose}}||^2} + \frac{\hat{\mathbf{a}}^{\text{T}} \mathbf{a}_i^{\text{pose}}}{||\mathbf{a}_i^{\text{pose}}||^2} \right) . \tag{6}$$

The model is then constructed by using these weights in (5) and finding the deformation modes from samples of the complete training set. Note that this decomposition does not guarantee orthogonality of shape or appearance modes, but we did not find this to be an issue in our application.

3.2. Local deformation modes

In this section we propose a method to obtain modes for local deformations such as eye blinking. This can be achieved by a modified version of the method described in

the previous section. Firstly shape and appearance modes which model blinking are learned from a video containing blinking with no other head motion. Directly applying the method in section 3.1 to remove these blinking modes from the training set introduces artifacts. The reason for this is apparent when considering the shape mode associated with blinking in which the majority of the movement is in the eyelid. This means that if the eyes are in a different position relative to the centroid of the face (for example if the mouth is open, lowering the centroid) then the eyelid is moved toward the mean eyelid position, even if this artificially opens or closes the eye. Instead of computing the weights of absolute coordinates in (6) we therefore propose to use relative shape coordinates using a Laplacian operator:

$$c_i^{\text{blink}} = \frac{1}{2}\left(\frac{L(\hat{\mathbf{s}})^{\mathrm{T}} L(\mathbf{s}_i^{\text{blink}})}{||L(\mathbf{s}_i^{\text{blink}})||^2} + \frac{\hat{\mathbf{a}}^T \mathbf{a}_i^{\text{blink}}}{||\mathbf{a}_i^{\text{blink}}||^2}\right). \quad (7)$$

The Laplacian operator $L()$ is defined on a shape sample such that the relative position, δ_i of each vertex i within the shape can be calculated from its original position $\mathbf{p}_i$ using

$$\delta_i = \sum_{j \in \mathcal{N}} \frac{\mathbf{p}_i - \mathbf{p}_j}{||d_{ij}||^2}, \quad (8)$$

where $\mathcal{N}$ is a one-neighborhood defined on the AAM mesh and d_{ij} is the distance between vertices i and j in the mean shape. This approach correctly normalizes the training samples for blinking, as relative motion within the eye is modeled instead of the position of the eye within the face.

3.3. Segmenting AAMs into regions

Different regions of the face can be moved nearly independently, a fact that has previously been exploited by segmenting the face into regions, which are modeled separately and blended at their boundaries [2, 9, 22]. While this approach tends to be followed in 3D models, it is difficult to apply to synthesizing with AAMs as these are not invariant to 3D pose, and mixing components could result in implausible instances where different regions have different pose.

The decomposition into pose and deformation components in (3) allows us to further separate the deformation components according to the local region they affect. We split the model into R regions and model its shape according to:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{K} c_i \mathbf{s}_i^{\text{pose}} + \sum_{j=1}^{R} \sum_{i \in I_j} c_i \mathbf{s}_i^j \ , \quad (9)$$

where I_j is the set of component indices associated with region j. The modes for each region are learned by only considering a subset of the model's vertices according to manually selected boundaries marked in the mean shape. Modes are iteratively included up to a maximum number, by greedily adding the mode corresponding to the region which allows the model to represent the greatest proportion of the observed variance in the training set. The analogous model is used for appearance. Linear blending is applied locally near the region boundaries.

We use this approach to split the face into an upper and lower half. The advantage of this is that changes in mouth shape during synthesis cannot lead to artifacts in the upper half of the face. Since global modes are used to model pose there is no risk of the upper and lower halves of the face having a different pose.

3.4. Extending the domain of an existing AAM

This section describes a method to extend the spatial domain of a previously trained AAM without affecting the existing model. In our case it was employed to extend a model that was trained only on the face region to include hair and ear regions in order to add more realism.

The set of N training images for the existing AAM is known, as are the original model coefficient vectors $\{\mathbf{c}_j\}_{j=1}^N$, $\mathbf{c}_j \in \mathcal{R}^M$ for these images. We proceed by labeling the regions to be included in the model, resulting in a new set of N training shapes $\{\tilde{\mathbf{s}}_j^{\text{ext}}\}_{j=1}^N$ and appearances $\{\tilde{\mathbf{a}}_j^{\text{ext}}\}_{j=1}^N$. Given the original model with M modes, the new shape modes, $\{\mathbf{s}_i\}_{i=1}^M$, should satisfy the following constraint

$$\left[\ \tilde{\mathbf{s}}_1^{\text{ext}} \dots \tilde{\mathbf{s}}_N^{\text{ext}}\ \right] = \left[\ \mathbf{s}_1 \dots \mathbf{s}_M\ \right] \left[\ \mathbf{c}_1 \dots \mathbf{c}_N\ \right], \quad (10)$$

which states that the new modes can be combined, using the original model coefficients, to reconstruct the extended training shapes $\tilde{\mathbf{s}}_j^{\text{ext}}$. Assuming that the number of training samples N is larger than the number of modes M the new shape modes can be obtained as the least-squares solution. New appearance modes are found analogously.

3.5. Adding regions with static texture

Since the teeth and tongue are occluded in many of the training examples, the synthesis of these regions contains significant artifacts when modeled using a standard AAM. To reduce these artifacts we use a fixed shape and texture for the upper and lower teeth. The displacements of these static textures are given by the displacement of a vertex at the center of the upper and lower teeth respectively. The teeth are rendered before the rest of the face, ensuring that the correct occlusions occur. A visual comparison is provided in figure 4(h).

4. Synthesis framework

Our synthesis model takes advantage of an existing TTS approach known as cluster adaptive training (CAT). The AAM described in the previous section is used to express

each frame in the training set as a low dimensional vector. The audio and video data are modeled using separate streams within a CAT model, a brief overview of which is given next.

4.1. Cluster adaptive training (CAT)

Cluster adaptive training (CAT) [28] is an extension to hidden Markov model text-to-speech (HMM-TTS). HMM-TTS is a parametric approach to speech synthesis [29] which models quinphones using HMMs with five emitting states. Concatenating the HMMs and sampling from them produces a set of parameters which can then be resynthesized into synthetic speech. Typically, a decision tree is used to cluster the quinphones to handle sparseness in the training data. For any given quinphone the means and variances to be used in the HMMs may be looked up using the decision tree.

The key addition of CAT is the use of multiple decision trees to capture speaker- or emotion-dependent information. Figure 2 shows the structure of the CAT model. Each cluster has its own decision tree and the means of the HMMs are determined by finding the mean for each cluster and combining them using the formula

$$\boldsymbol{\mu}_m^{\text{expr}} = \mathbf{M}_m \boldsymbol{\lambda}^{\text{expr}}, \tag{11}$$

where $\boldsymbol{\mu}_m^{\text{expr}}$ is the mean for a given expression, m is the state of the HMM, $\mathbf{M}_m$ is the matrix formed by combining the means from each cluster and $\boldsymbol{\lambda}^{\text{expr}}$ is a weight vector.

Each cluster in CAT may be interpreted as a basis defining an expression space. To form the bases, each cluster is initialized using the data of one emotion (by setting the $\boldsymbol{\lambda}$'s to zero or one as appropriate). The Maximum-Likelihood criterion is used to update all the parameters in the model (weights, means and variances, and decision trees) iteratively. The resulting $\boldsymbol{\lambda}$'s may interpreted as coordinates within the expression space. By interpolating between $\boldsymbol{\lambda}^{\text{expr}_1}$ and $\boldsymbol{\lambda}^{\text{expr}_2}$ we can synthesize speech with an expression between two of the originally recorded expressions. Since the space is continuous it is possible to synthesize at any point in the space and generate new expressions. For more details the reader is referred to [15].

5. Experiments

We collected a corpus of 6925 sentences, divided between 6 emotions; neutral, tender, angry, afraid, happy and sad. From the data 300 sentences were held out as a test set and the remaining data was used to train the speech model. The speech data was parameterized using a standard feature set consisting of 45 dimensional Mel-frequency cepstral coefficients, log-F0 (pitch) and 25 band aperiodicities, together with the first and second time derivatives of these features. The visual data was parameterized using

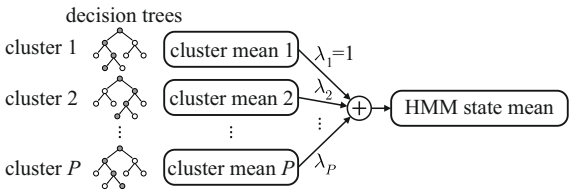


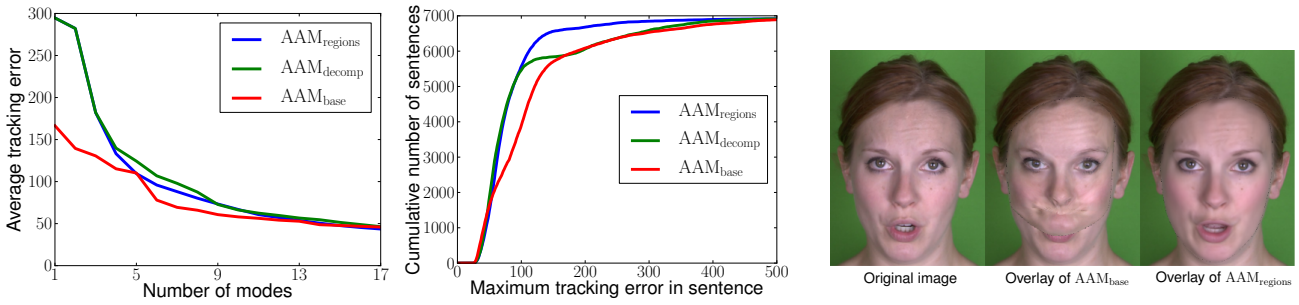
Figure 2: Cluster adaptive training (CAT). *Each cluster is represented by a decision tree and defines a basis in expression space. Given a position in this expression space defined by* $\boldsymbol{\lambda}^{\text{expr}} = [\lambda_1 \dots \lambda_P]$ *the properties of the HMMs to use for synthesis can be found as a linear sum of the cluster properties.*

the AAMs described below. We trained different AAMs in order to evaluate the improvements obtained with the proposed extensions. In each case the AAM was controlled by 17 parameters and the parameter values and their first time derivatives were used in the CAT model.

The first model used, AAM_{base}, is a standard AAM as described in [7], built from 71 training images in which 47 facial keypoints were labeled by hand. Additionally, contours around both eyes, the inner and outer lips, and the edge of the face were labeled and points were sampled at uniform intervals along their length. The second model, $\text{AAM}_{\text{decomp}}$, separates both 3D head rotation (modeled by two modes) and blinking (modeled by one mode) from the deformation modes as described in sections 3.1 and 3.2. The third model, $\text{AAM}_{\text{regions}}$, is built in the same way as $\text{AAM}_{\text{decomp}}$ expect that 8 modes are used to model the lower half of the face and 6 to model the upper half, see section 3.3. The final model, AAM_{full}, is identical to $\text{AAM}_{\text{regions}}$ except for the mouth region which is modified as described in section 3.5. Please see the supplementary video for samples of synthesis.

5.1. Evaluating AAM reconstruction

In the first experiment we quantitatively evaluate the reconstruction error of each AAM on the complete data set of 6925 sentences which contains approximately 1 million frames. The reconstruction error was measured as the L_2 norm of the per-pixel difference between an input image warped onto the mean shape of each AAM and the generated appearance. Figure 3(a) shows how reconstruction errors vary with the number of AAM modes. It can be seen that while with few modes, AAM_{base} has the lowest reconstruction error, as the number of modes increases the difference in error decreases. In other words, the flexibility that semantically meaningful modes provide does not come at the expense of reduced tracking accuracy. In fact we found the modified models to be more robust than the base model, having a lower worst case error on average, as shown in figure 3(b). This is likely due to $\text{AAM}_{\text{regions}}$ and $\text{AAM}_{\text{decomp}}$ being better able to generalize to unseen examples as they



(a) (b) (c)

Figure 3: AAM reconstruction errors. *(a) Average errors* vs. *number of AAM modes. It can be seen that the average errors of all models converge as the number of modes increases. (b) Cumulative number of sentences remaining below a given tracking error (for models using 17 modes). It can be seen that the proposed AAM extensions of* $\mathrm{AAM}_{\mathrm{regions}}$ *and* $\mathrm{AAM}_{\mathrm{decomp}}$ *reduce the maximum errors compared to the standard* $\mathrm{AAM}_{\mathrm{base}}$*. (c) An example of tracking failure for* $\mathrm{AAM}_{\mathrm{base}}$ *since this combination of mouth shape and expression did not appear in the training set.*

do not overfit the training data by learning spurious correlations between different face regions. An example where this causes $\mathrm{AAM}_{\mathrm{base}}$ to fail is given in figure 3(c).

5.2. User studies

We carried out a number of large-scale user studies in order to evaluate the perceptual quality of the synthesized videos. The experiments were distributed via a crowd sourcing website, presenting users with videos generated by the proposed system.

5.2.1 Preference studies

To determine the qualitative effect of the choice of AAM on the final system we carried out preference tests on systems built using the different AAMs. For each preference test 10 sentences in each of the six emotions were generated with two models rendered side by side. Each pair of AAMs was evaluated by 10 users who were asked to select between the left model, right model or having no preference (the order of our model renderings was switched between experiments to avoid bias), resulting in a total of 600 pairwise comparisons per preference test. In this experiment the videos were shown without audio in order to focus on the quality of the face model.

From table 1 it can be seen that $\mathrm{AAM}_{\mathrm{full}}$ achieved the highest score, and that $\mathrm{AAM}_{\mathrm{regions}}$ is also preferred over the standard AAM. This preference is most pronounced for expressions such as *angry*, where there is a large amount of head motion and less so for emotions such as *neutral* and *tender* which do not involve significant movement of the head. This demonstrates that the proposed extensions are particularly beneficial to expressive VTTS.

5.2.2 Comparison with other talking heads

In order to compare the output of different VTTS systems users were asked to rate the realism of sample synthesized sentences on a scale of 1 to 5, with 5 corresponding to 'completely real' and 1 to 'completely unreal'. Sample sentences that were publicly available were chosen for the evaluation, and scaled to a face region height of approximately 200 pixels. The degree of expressiveness of the systems range from neutral speech only to highly expressive. The results in Table 2 show that the system by Liu *et al.* was rated most realistic among the systems for neutral speech and with a small degree of expressiveness. The proposed system performs comparably to other methods in the neutral speech category, while for larger ranges of expression it achieved a significantly higher score than the system by Cao *et al.* In this study each system was rated by 100 users.

5.2.3 Emotion recognition study

In the final study we evaluated the ability of the proposed VTTS system to express a range of emotions. Users were presented either with video or audio clips of a single sentence from the test set and were asked to identify the emotion expressed by the speaker, selecting from a list of six emotions. The synthetic video data for this evaluation was generated using the $\mathrm{AAM}_{\mathrm{regions}}$ model. We also compared with versions of synthetic video only and synthetic audio only, as well as cropped versions of the actual video footage. In each case 10 sentences in each of the six emotions were evaluated by 20 people, resulting in a total sample size of 1200. Example frames showing each emotion are given in figure 4.

The average recognition rates are 73% for the captured footage, 77% for our generated video (with audio), 52% for the synthetic video only and 68% for the synthetic audio only. These results indicate that the recognition rates

AAM base	AAM decomp	AAM region	AAM full	Orig. video	No pref.
36	**37**				28
34		**48**			18
34			**53**		13
	35	**39**			25
	33	**51**			16
		30	**50**		20
			14	**82**	4

AAM base	AAM decomp	AAM region	AAM full	Orig. video	No pref.
40	36				24
37		**47**			15
45			41		14
	42	35			22
	38	**48**			13
		28	**46**		26
			11	**83**	6

AAM base	AAM decomp	AAM region	AAM full	Orig. video	No pref.
33	**45**				22
20		**75**			5
22			**72**		6
	29	**52**			19
	31	**49**			19
		33	**43**		24
			12	**84**	4

Table 1: Pairwise preference tests *between different models. Scores shown as percentages of all votes for:* **(left)** *all emotions,* **(middle)** *neutral, and* **(right)** *angry. There is a preference for the refined models for the average score over all emotions, this is mostly due to the emotions with a large amount of movement, such as angry. The preference for the proposed model over other AAMs is particularly clear for emotions with significant head motion, such as angry shown in the right table.*

Method	**Expressions**	**Realism Score**	
Chang and Ezzat [6]	neutral	3.3	(4.5)
Deena *et al.* [10]	neutral	3.4	(3.7)
Wang *et al.* [26] *male*	neutral	4.0	
Wang *et al.* [26] *fem.*	neutral	3.9	
Liu *et al.* [16]	neutral	4.3	(4.6)
this paper	neutral	3.7	(4.4)
Liu *et al.* [16]	small range	3.6	
Melenchon *et al.* [18]	small range	3.1	
Cao *et al.* [5]	small range	2.6	
Cao *et al.* [5]	large range	2.7	
this paper	large range	3.8	(4.4)

Table 2: Comparative user study. *Users rated the realism of sample sentences generated using different VTTS systems where higher values correspond to more realistic output. Scores for actual footage are shown in the last column for systems where data was available. It can be seen that for high expressiveness the proposed system achieves a higher score than that by Cao* et al.

for synthetically generated results are comparable, or even slightly higher than for the real footage. This may be due to the stylization of the expression in the synthesis. Confusion matrices between the different expressions are shown in figure 5. Tender and neutral expressions are most easily confused in all cases. While some emotions are better recognized from audio only, the overall recognition rate is higher when using both cues.

6. Conclusions and future work

In this paper we have demonstrated a complete visual text-to-speech system which is capable of creating near-videorealistic synthesis of expressive text. We have carried out user studies showing that its performance is state of the art by comparing directly to other current VTTS systems. To improve performance of our system we have adapted active appearance models to reduce the main artifacts resulting from using a person specific active appearance model for rendering. In the future we plan to extend the system so that the identity of the speaker is controllable as well as their expression.

Acknowledgments. We are grateful to all researchers in the Speech Technology Group at Toshiba Research Europe for their work on the speech synthesis side of the model. We also thank Oliver Woodford, Sam Johnson and Frank Perbet for helpful discussions on the paper.

References

[1] I. Albrecht, M. Schröder, J. Haber, and H. Seidel. Mixed feelings: expression of non-basic emotions in a muscle-based talking head. *Virt. Real.*, 8(4):201–212, 2005. 1, 2

[2] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. *SIGGRAPH*, pages 187–194, 1999. 4

[3] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer. High resolution passive facial performance capture. *ACM TOG*, 29(4), 2010. 2

[4] M. Brand. Voice puppetry. In *SIGGRAPH*, pages 21–28, 1999. 2

[5] Y. Cao, W. Tien, P. Faloutsos, and F. Pighin. Expressive speech-driven facial animation. *ACM TOG*, 24(4):1283–1302, 2005. 1, 2, 7

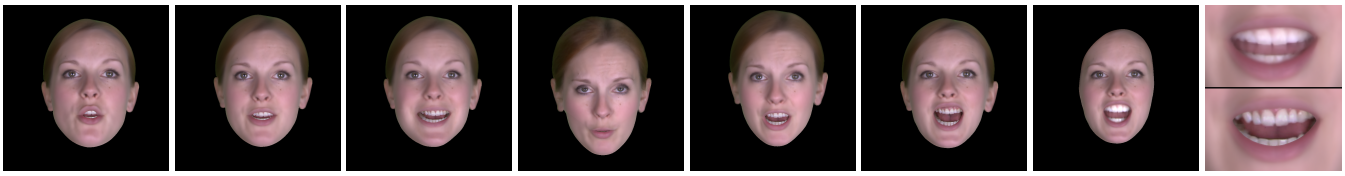
[6] Y. Chang and T. Ezzat. Transferable videorealistic speech animation. In *SIGGRAPH*, pages 143–151, 2005. 1, 2, 7

[7] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE PAMI*, 23(6):681–685, 2001. 1, 2, 3, 5

[8] D. Cosker, S. Paddock, D. Marshall, P. Rosin, and S. Rushton. Towards perceptually realistic talking heads: models, methods and mcgurk. In *Symp. Applied perception in graphics and visualization*, pages 151–157, 2004. 2

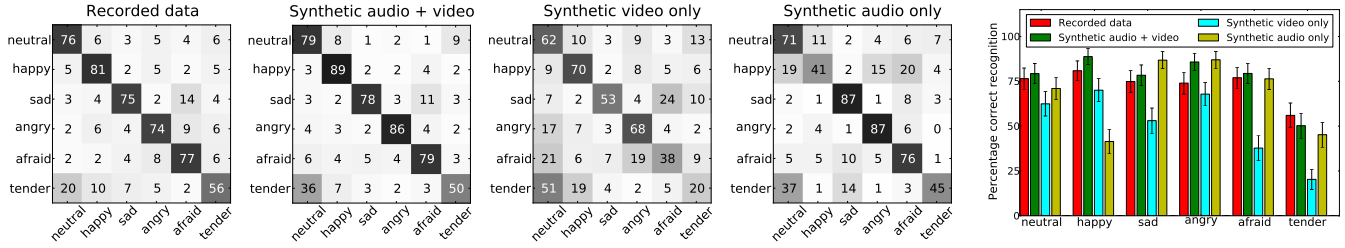
[9] F. De la Torre and M. Black. Robust parameterized component analysis: theory and applications to 2d facial appearance models. *Computer Vision and Image Understanding*, 91(1):53–71, 2003. 2, 4

[10] S. Deena, S. Hou, and A. Galata. Visual speech synthesis by modelling coarticulation dynamics using a non-parametric



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 4: Example synthesis *for (a) neutral, (b) tender, (c) happy, (d) sad, (e) afraid and (f) angry, (g) same angry frame without teeth modifications or hair, (h) close up of teeth, (top) before modification and (bottom) after.*



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

Figure 5: Emotion recognition *for (a) real video cropped to face, (b) synthetic audio and video, (c) synthetic video only and (d) synthetic audio only. In each case 10 sentences in each emotion were evaluated by 20 different people. (e) gives the recognition rate for each emotion along with the 95% confidence interval.*

switching state-space model. In *ICMI-MLMI*, pages 1–8, 2010. 1, 2, 7

[11] G. Edwards, A. Lanitis, C. Taylor, and T. Cootes. Statistical models of face images - improving specificity. *Image and Vision Computing*, 16(3):203–211, 1998. 2, 3

[12] T. Ezzat and T. Poggio. Miketalk: A talking facial display based on morphing visemes. In *In Proceedings of the Computer Animation Conference*, pages 96–102, 1998. 2

[13] X. Gao, Y. Su, X. Li, and D. Tao. A review of active appearance models. *IEEE Transactions on Systems, Man, and Cybernetics*, 40(2):145–158, 2010. 2

[14] J. Gonzalez-Mora, F. De la Torre, R. Murthi, N. Guil, and E. Zapata. Bilinear active appearance models. *ICCV*, pages 1–8, 2007. 2

[15] J. Latorre, V. Wan, M. J. F. Gales, L. Chen, K. Chin, K. Knill, and M. Akamine. Speech factorization for HMM-TTS based on cluster adaptive training. In *Interspeech*, 2012. 5

[16] K. Liu and J. Ostermann. Realistic facial expression synthesis for an image-based talking head. In *International Conference on Multimedia & Expo*, pages 1–6, 2011. 1, 2, 7

[17] W. Ma, A. Jones, J. Chiang, T. Hawkins, S. Frederiksen, P. Peers, M. Vukovic, M. Ouhyoung, and P. Debevec. Facial performance synthesis using deformation-driven polynomial displacement maps. *SIGGRAPH*, 27(5), 2008. 2

[18] J. Melenchón, E. Martínez, F. De la Torre, and J. Montero. Emphatic visual speech synthesis. *Trans. Audio, Speech and Lang. Proc.*, 17(3):459–468, 2009. 7

[19] I. Pandzic, J. Ostermann, and D. Millen. User evaluation: synthetic talking faces for interactive services. *The Visual Computer*, 15(7):330–340, 1999. 1

[20] E. Sifakis, A. Selle, A. Robinson-Mosher, and R. Fedkiw. Simulating speech with a physics-based facial muscle model. In *SCA ACM/Eurographics*, pages 261–270, 2006. 2

[21] S. Taylor, M. Mahler, B. Theobald, and I. Matthews. Dynamic units of visual speech. In *Eurographics Symposium on Computer Animation*, pages 275–284, 2012. 2

[22] J. Tena, F. De la Torre, and I. Matthews. Interactive region-based linear 3d face models. *ACM TOG*, 30(4):76, 2011. 4

[23] B. Theobald, J. Bangham, I. Matthews, and G. Cawley. Near-videorealistic synthetic talking faces: implementation and evaluation. *Speech Comm.*, 44(14):127–140, 2004. 1, 2

[24] K. Wampler, D. Sasaki, L. Zhang, and Z. Popović. Dynamic, expressive speech animation from a single mesh. In *SCA ACM/Eurographics*, pages 53–62, 2007. 2

[25] L. Wang, W. Han, X. Qian, and F. Soong. Photo-real lips synthesis with trajectory-guided sample selection. In *Speech Synth. Workshop, Int. Speech Comm. Assoc.*, 2010. 1, 2

[26] L. Wang, W. Han, F. Soong, and Q. Huo. Text driven 3D photo-realistic talking head. In *Interspeech*, pages 3307–3308, 2011. 2, 7

[27] K. Waters and T. Levergood. DECface: A system for synthetic face aplications. *Multimedia Tools and Applications*, 1(4):349–366, 1995. 1

[28] H. Zen, N. Braunschweiler, S. Buchholz, M. Gales, K. Knill, S. Krstulović, and J. Latorre. Statistical Parametric Speech Synthesis Based on Speaker and Language Factorization. *IEEE Trans. Audio Speech Lang. Process.*, 20(5), 2012. 5

[29] H. Zen, K. Tokuda, and A. Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1154, November 2009. 1, 5