

Active Appearance Models for Photorealistic Visual Speech Synthesis

Wesley Mattheyses, Lukas Latacz and Werner Verhelst

Vrije Universiteit Brussel, Dept. ETRO-DSSP,
Interdisciplinary Institute for Broadband Technology IBBT, Brussels, Belgium

{wmatthey, llatacz, wverhels}@etro.vub.ac.be

Abstract

The perceived quality of a synthetic visual speech signal greatly depends on the smoothness of the presented visual articulators. This paper explains how concatenative visual speech synthesis systems can apply active appearance models to achieve a smooth and natural visual output speech. By modeling the visual speech contained in the system's speech database, a diversification between the synthesis of the shape and the texture of the talking head is feasible. This allows the system to accurately balance between the articulation strength of the visual articulators and the signal smoothness of the visual mode in order to optimize the synthesis. To improve the synthesis quality, an automatic database normalization strategy has been designed that removes variations from the database which are not related to speech production. As was verified by a perception experiment, this normalization strategy significantly improves the perceived signal quality.

Index Terms: audiovisual speech synthesis, AAM modeling

1. Introduction

Audiovisual text-to-speech (AVTTS) systems generate an audiovisual speech signal based on a written input text. Their functionality can be applied in various domains where machine-human communication is needed, like in e-commerce and e-learning environments. In comparison to classical auditory-only speech synthesis, the addition of a virtual talking head to the artificial speech is advantageous since it will make the user feel more confident and attentive [1] and will consequently enhance the quality of the communication. In previous work we have designed a multimodal speech synthesis system that can create a synthetic audiovisual speech signal by concatenating audiovisual speech segments, selected from a multimodal speech database [2]. One of the challenges in concatenative (audio-)visual speech synthesis is achieving a smooth and natural visual speech signal. Any change in appearance of the lips, teeth or other visual articulators that is unlike natural speech will be easily noticed by a viewer and will therefore decrease the perceived naturalness. In this paper we will elaborate on the use of active appearance models in order to achieve a high quality visual speech signal. We will explain how their properties can be exploited to successfully generate a smooth and natural sequence of mouth images which vary in accordance with the target speech.

2. Active Appearance Modeling

2.1. Problem Statement

Traditionally, for data-based 2D (audio-)visual speech synthesis (e.g., [2][3]) the information contained in the visual speech

database is treated as sequences of static images. To achieve quality speech synthesis, an accurate analysis of this visual speech data is required since the system needs to carefully select appropriate segments from the database and it needs to optimize the video joins in order to create a smooth output signal. However, processing the data as static images makes it very hard to differentiate between aspects concerning the speech movements (e.g., lip and tongue movements) and aspects concerning the overall appearance of the mouth area (e.g., visibility of the teeth, colors, shadows, etc.). For instance, in [2] image morphing is applied to optimize the concatenation of two video segments. A careful inspection of this system's synthetic audiovisual speech showed that a strong concatenation smoothing is necessary to avoid sudden steep changes in the overall appearance of the mouth area. On the other hand, only a minor smoothing can be applied in order to avoid under-articulation effects: a stronger smoothing of the lip movements results in visual speech that appears 'mumbled', since the mouth movements are too slow and too limited to match with the stronger articulations present in the accompanying auditory speech. This indicates that there is the necessity for a technique which makes it possible to differentiate between the different aspects of the visual speech information, like lip movements, teeth and tongue visibility, illumination changes, etc. Such a differentiation is useful to balance the concatenation smoothing strength and it will also provide additional meta-data concerning the visual speech which can be used by the segment selection algorithms.

2.2. Active Appearance Models

2D active appearance models (AAMs) [4] are statistical models that are able to project a set of similar images into a model-space. After projection on the model, the images are represented by their corresponding model parameters. In addition, a trained AAM makes it possible to generate a new image from a set of AAM model parameters that is given as input. AAMs model two different aspects of an image: the shape and the texture. The shape of an image is defined by a set of landmark points that indicate the position of certain objects that are present in each training image. To train an AAM, this shape has to be defined manually for each training image. The texture of an image is determined by its pixel values, which are sampled over triangles defined by the image's landmark points. This texture is sampled using the shape-normalized equivalent of the image: before sampling the triangles, the image is warped by matching its landmark points on the mean shape of the AAM (i.e., the mean value of every landmark point sampled over the training images). Thus, in order to project an image on an AAM, the image is defined by a vector containing the landmark positions, i.e. its shape S , and a vector containing the pixel values of its shape-normalized equivalent, i.e. its texture

T . From all training shapes S_i , the mean shape S_m is calculated and a PCA calculation is performed on the normalized shapes $S_i - S_m$, resulting in a set of eigenshapes P_s which determine the AAM shape-model. Likewise, the mean texture T_m and the AAM texture-model (determined by eigentextures P_t) is calculated from all training textures. After training the AAM, any image with shape S and texture T can be projected on the AAM by searching iteratively for the most appropriate model parameters (shape-parameters B_s and texture-parameters B_t) to reconstruct the original shape and texture using the shape- and texture-model:

$$S_{recon} = S_m + P_s \times B_s, \quad T_{recon} = T_m + P_t \times B_t \quad (1)$$

After projection on the AAM, the image is represented by its shape and its texture parameters. Furthermore, from an unseen set of shape and texture parameters and a trained AAM, a new shape S^{new} and a new texture T^{new} can be calculated using Eq.1. From these a new image can be generated by warping the shape-normalized texture T^{new} (aligned with the mean shape S_m) towards the new shape S^{new} . Note that the shape of an image only needs to be determined manually during the training phase. Once the AAM has been trained, the shape of an image (i.e., its landmarks) can be determined automatically by projecting the image on the AAM and by calculating its shape from the computed shape-parameters. For more details on the iterative model-search which is necessary to project an image on the model, the reader is referred to [4].

3. Visual Speech Synthesis Using AAMs

3.1. Introduction

AAMs are used to represent image data by means of shape and texture parameters. As was explained in section 2.1, the ability to differentiate between shape and texture properties makes AAMs very suited for visual speech synthesis purposes. When the visual speech information, contained in the database of the AVTTS system, has been transformed into trajectories of AAM parameters (by mapping each frame on a set of shape and texture parameters), visual speech synthesis can be achieved by selecting and concatenating the appropriate sets of sub-trajectories from the database. From these new concatenated trajectories the final output video can be created by generating the output frames from parameter values sampled from these trajectories. In [5] AAMs are used to model the complete face of a speaker. However, in order to achieve a maximal lip-readability, we opted to build an AAM which only models the mouth area of a talking head. This way, all variance captured by the AAM originates from variations of the lips, teeth, tongue, etc. In addition, this work aims to exploit the ability of AAMs to differentiate between shape and texture information as much as possible. We investigated on several techniques to benefit from the fact that the visual speech information is no longer represented by static frames but by two discrete sets of time-varying model parameters.

3.2. Database Preparation

A first step towards AAM-based speech synthesis consists in building the appropriate AAM that is able to model the data from the AVTTS system's visual speech database. For the work described in this paper, the LIPS2008 audiovisual speech database [6] has been used. We designed an iterative technique to build a high quality AAM that preserves much image detail, while the amount of manual work is limited. First, the mouth



Figure 1: *From left to right: original frame, its landmarking (denoting the shape) and the AAM reconstructed image*

area of 20 random frames from the database was landmarked manually, indicating the position of the lips, cheeks, chin and nose as shown in Fig.1. From these frames and their landmarking, a first AAM was build using standard AAM algorithms [7]. Afterwards, this trained AAM was used to track 100 random sentences contained in the database, i.e. to calculate their shape and texture parameters (and the corresponding landmarks) as was explained in section 2.2. A clustering on the calculated parameters was performed to determine 50 visually distant frames, of which the shape was re-labeled manually. These frames were used to train an improved AAM, which was then applied to automatically track the whole database. A clustering on this new data resulted in 160 images that were selected as final training set. Their automatically determined landmarking was checked manually and corrected if necessary, after which the final AAM was build from this data. The AAM was designed to retain 97% of the variation contained in the training set, resulting in 8 eigenvectors that represent the shape model and 134 eigenvectors that represent the texture model. Finally, this AAM was used to transform the whole visual database into sequences of shape and texture parameters. An example of an original frame extracted from the database video, its automatic landmarking and its AAM reconstruction using its shape and texture parameters is given in Fig.1.

3.3. Visual Synthesis

The basic concept of our audiovisual speech synthesis strategy has already been described in [2]: the system selects audiovisual segments from the audiovisual speech database, containing a natural combination of audio and video to ensure a maximal coherence between the two output speech modes. This strategy is based on the unit-selection technique: to synthesize a target sentence, the system searches in the database for audiovisual segments using target costs and join costs. Since the visual speech database has been projected on the AAM, the video segments that are selected by the unit-selection can be represented by their corresponding sub-trajectories of model parameters. This creates the opportunity to accurately smooth the video concatenations by overlapping and interpolating the AAM parameters of the frames at the boundaries of the video segments. Eq.2 illustrates this for the concatenation of two video segments, both represented by a series of vectors containing the model parameters, $(\mathbf{B}_1^1, \mathbf{B}_2^1, \dots, \mathbf{B}_m^1)$ and $(\mathbf{B}_1^2, \mathbf{B}_2^2, \dots, \mathbf{B}_n^2)$, with resulting joined video signal $(\mathbf{B}_1^j, \mathbf{B}_2^j, \dots, \mathbf{B}_{m+n-1}^j)$. In Eq.2, the parameter S determines the smoothing strength: a larger value of S will cause a more pronounced interpolation at the concatenation points. The major benefit of the AAM-based synthesis approach is that the amount of smoothing can be diversified between the shape and the texture trajectories: a light smoothing can be applied to the shape parameters to avoid visual under-articulation, while a stronger smoothing is applied to the texture parameters

to ensure a visually smooth output signal.

Overlap: $B_m^j = 0.5 \times (B_m^1 + B_1^2)$

Interpolation: For $1 \leq k \leq m+n-1$: $B_k^j =$

$$\begin{cases} B_k^1 & 1 \leq k < m-S \\ \frac{m-k}{S+1} B_k^1 + \frac{(S+1)-(m-k)}{S+1} B_m^j & m-S \leq k < m \\ \frac{(S+1)-(k-m)}{S+1} B_m^j + \frac{k-m}{S+1} B_{k-m+1}^2 & m < k \leq m+S \\ B_{k-m+1}^2 & m+S < k \leq m+n-1 \end{cases} \quad (2)$$

After concatenation, a single trajectory for each model parameter is obtained. The target output video containing the mouth-area of the talking face is created by generating the video frames from these trajectories using the AAM. An overlay of this video signal on a background video containing the other parts of the face creates the final visual output speech.

4. Database Normalization

4.1. Introduction

The quality of data-based speech synthesis depends strongly on the properties of the speech database used. When registering an (audio-)visual speech database, it is impossible to retain exactly the same recording conditions throughout the whole database. For instance, the LIPS2008 database contains some slight changes of the head position of the speaker, together with small variations of the illumination conditions and some color shifts. Although these variations are subtle, they can cause serious concatenation artifacts: since these features are not correlated with the speech, while synthesizing they will be randomly selected and concatenated. Here we propose a technique to minimize the effect of such variations present in the database by exploiting the properties of AAMs: a video frame is represented by its shape and its texture through its shape and texture parameters, respectively. These model parameters are each linked to an eigenvector, resulting from PCA calculations on the shapes and the textures contained in the training set (see section 2.2). We found that in practice, many of these eigenvectors can be linked to a certain physical property. For example, the first shape parameter of our trained AAM influences the amount of mouth-opening while the second shape parameter influences the head rotation. Likewise, the first texture parameter affects the appearance of shadows on the face, while the second texture parameter involves the presence of teeth in the image. These properties can be used to reduce the undesired database variations: the model parameters that do not represent changes that are correlated with the speech should be kept constant. An appropriate normalization value is zero, since all-zero model parameters lead to the mean AAM image (see Eq.1).

4.2. Normalization Strategy

A normalized version of the AAM-projected visual speech database has been created in which for every frame all non-speech related model parameters are set to zero in order to minimize the variations which are not related to speech production. To determine which parameters to normalize, two different measures have been designed. A first measure is based on the assumption that the visual representations of distinct instances of a same phoneme will look similar (since this is more valid for some phonemes than for others, we process all different phonemes that are present in the database and the mean measure among these phonemes will be used as will be explained later). This implies that when a parameter is sufficiently correlated with the speech, its values measured at several database instances of the

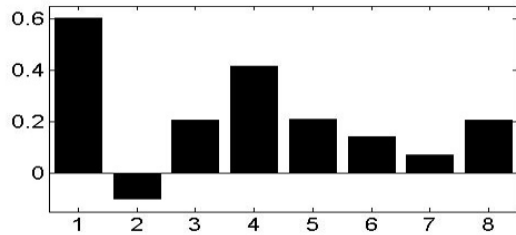


Figure 2: D_j^{var} values calculated for the eight shape-parameters

same phoneme will be more or less equal (some variation will exist due to visual co-articulation, etc.). Therefore, for every phoneme we selected 50% of its database occurrences and at the middle of each of these instances we sampled the shape and texture parameters. The mean M and the variation S of this data were calculated, resulting in values M_{ij} and S_{ij} where index i corresponds to the different phonemes present in the database and index j corresponds to the different model parameters. Then, for each phoneme, we selected a certain amount of random frames from the database. The model parameters of these frames were sampled, after which the mean M_{ij}^{rand} and variance S_{ij}^{rand} of this random set of parameters were calculated. The amount of random samples measured for a certain phoneme was the same as the amount of instances that were used to calculate M_{ij} and S_{ij} for that particular phoneme. Then, the relative difference between S_{ij} and S_{ij}^{rand} was calculated:

$$D_{ij}^{var} = (S_{ij}^{rand} - S_{ij}) / S_{ij}^{rand} \quad (3)$$

Finally, a single measure for each parameter (D_j^{var}) was acquired by taking the mean of D_{ij}^{var} among all phonemes (i.e. over index i). These values express the relative difference between the intra-phoneme variation and the overall variation of a parameter, and should be large for speech-correlated parameters. Fig.2 shows the D_j^{var} values that were calculated for the shape-parameters of our trained AAM. Another approach that has been applied to determine the speech-correlated parameters is to first resynthesize some random sentences from the database (these sentences are excluded from the database to avoid them to be selected by the unit-selection). Then, the parameter trajectories of these synthesized sentences are synchronized with the trajectories of the original sentences using the phonetic segmentation of the original and synthesized versions. For each sentence (index n) and for each parameter (index j), the Euclidean differences D_{nj}^{syn} between the original and synthesized trajectories are measured. Since an original trajectory's mean and variation vary a lot among the model parameters, every original trajectory OT_{nj} is first scaled to unit variance and zero mean. Consecutively, the mean and variance of the corresponding synthesized trajectory ST_{nj} are scaled using the mean and the variance of OT_{nj} . This way, a minimal distance between OT_{nj} and ST_{nj} is measured when they are similar in both mean, variation and shape. For every sentence (i.e. a fixed value of index n), the measured differences D_{nj}^{syn} are scaled between zero and one to cancel out the global synthesis quality of the sentence. Finally, calculating the mean among all sentences results in a single value D_j^{syn} for each parameter. This value will be larger for parameters which are not correlated with the speech: the values D_{nj}^{syn} are calculated by comparing the model parameters of video frames belonging to two different database



Figure 3: Reconstruction of a database frame using the original model parameters (left) and the normalized model parameters (right)

instances of the same phoneme. By constructing these comparison pairs using speech synthesis, we ensure that the two phoneme instances are similar in terms of visual context, linguistic properties, etc. which implies that their visual representations will be much alike. Finally, to determine which parameters to normalize, measures D_j^{var} and D_j^{syn} are combined. First, for both measures the 30% shape/texture parameters least correlated with the speech are selected. Then, from this selection a final set is determined as the parameters that were selected by both measures, augmented with the parameters that were selected by only one measure and which represent less than 1% model variation (i.e., the parameter's corresponding eigenvector holds less than 1% of the variation contained in the training set that was used to build the AAM). For our AAM, this resulted in the selection of 1 shape-parameter and 35 texture parameters for normalization. An example of an AAM reconstructed frame before and after normalizing the model parameters is shown in Fig.3.

4.3. Evaluation

To assess the effect of the database normalization on the synthesis quality, a perception experiment has been conducted. For this experiment, 15 audiovisual sentences were synthesized using both the original and the normalized version of the AAM-projected database. These samples were shown pairwise to the participants, who were asked to write down their opinion on the naturalness of the mouth area. They were instructed to pay attention to the smoothness and naturalness of the mouth movements/appearances and to the coherence between the auditory and visual speech. A 5-point comparative MOS scale [-2,2] was used to express their preference for the first or for the second sample of each pair. 9 participants joined the experiment, 1 of them aged above 50 and the others aged between 21-30. 7 of them can be considered speech experts. The results of the test are given in table 1; to construct this table, each sample was assigned a score [-2,2], after which a paired-sample Wilcoxon signed rank test was performed to test whether there exists a significant difference between the samples created using the original AAM parameters and the samples created using the normalized model parameters. The Z value of -8.26 shows that the normalization does improve the synthesis quality by removing slight head rotations and other variations in the database which are not correlated with the speech.

5. Discussion

One of the major challenges in data-based visual speech synthesis is to balance between signal smoothness and articulation strength. In this paper we have explained why AAMs are suited to tackle this problem and we have described an iterative tech-

Table 1: Listening test on the parameter normalization

Answers	
Total	135
Normalized < Original	5
Normalized = Original	40
Normalized > Original	90
Wilcoxon Test Statistics	
Z	-8.265
Significance	< 0.01

nique to build an AAM which can be used to transform the information contained in the visual speech database into a set of parameter trajectories. Once the AVTTS system's unit selection algorithms have selected an appropriate set of video segments, sub-trajectories can be extracted from the database. The ability of AAMs to separately model the shape and the texture information makes it possible to accurately fine-tune the concatenation of these sub-trajectories: the overall appearance can be easily smoothed to create an overall smooth signal, while the movements of the lips and the other visual articulators are still sufficiently pronounced (to avoid visual under-articulation). In addition, we proposed a database normalization strategy which makes it possible to remove many of the non-speech related variations present in the database. As has been verified by a perception experiment, this strategy significantly improves the perceived synthesis quality since it enhances the smoothness and naturalness of the synthetic visual speech. Sample syntheses using the normalized database can be found at <http://www.etro.vub.ac.be/Research/DSSP/DEMO/AVTTS/>.

6. Acknowledgments

The research reported on in this paper was partly supported by a research grant from the Faculty of Engineering Science, Vrije Universiteit Brussel.

7. References

- [1] Pandzic, I., Ostermann J. and Millen D., "Users Evaluation: Synthetic talking faces for interactive services", The Visual Computer, Volume 15 2330-2340, 1999
- [2] Mattheyses, W., Latacz, L. and Verhelst, W., "On the importance of audiovisual coherence for the perceived quality of synthesized visual speech", EURASIP Journal on Audio, Speech, and Music Processing, SI: Animating Virtual Speakers or Singers from Audio: Lip-Synching Facial Animation, 2009
- [3] Cosatto, E and Graf, H.P., "Photo-realistic talking-heads from image samples", IEEE Transactions on multimedia, Volume 2 152-163, 2000
- [4] Edwards, G.J., Taylor, C.J. and Cootes, T.F., "Interpreting Face Images using Active Appearance Models", Int. Conf. on Face and Gesture Recognition, 300-305, 1998
- [5] Theobald, B.J., Bangham, J.A., Matthews, I.A. and Cawley, G.C., "Near-videorealistic synthetic talking faces: implementation and evaluation", Speech Communication, Volume 44 127-140, 2004
- [6] Theobald, B.-J., Fagel, S., Bailly, G. and Elisei, F., "LIPS2008: Visual speech synthesis challenge", Interspeech 2008, 1875-1878, 2008
- [7] Stegmann, M.B., Ersbll, B.K., Larsen, R., "FAME - A Flexible Appearance Modelling Environment", IEEE Transactions on Medical Imaging, Volume 22(10), 1319-1331, 2003