

Optimized Photorealistic Audiovisual Speech Synthesis Using Active Appearance Modeling

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

Active appearance models can represent image information in terms of shape and texture parameters. This paper explains why this makes them highly suitable for data-based 2D audiovisual text-to-speech synthesis. We elaborate on how the differentiation between shape and texture information can be fully exploited to create appropriate unit-selection costs and to enhance the video concatenations. The latter is very important since for the synthetic visual speech a careful balancing between signal smoothness and articulation strength is required. Several optimization strategies to enhance the quality of the synthetic visual speech are proposed. By measuring the properties of each model parameter, an effective normalization of the visual speech database is feasible. In addition, the visual joins can be optimized by a parameter-specific concatenation smoothing. To further enhance the naturalness of the synthetic speech, a spectrum-based smoothing approach is introduced.

Index Terms: audiovisual speech synthesis, AAM modeling

1. Introduction

The purpose of an audiovisual text-to-speech (AVTTS) system is to generate a synthetic audiovisual speech signal based on a written input text. Many applications for these systems are imaginable, for instance in e-commerce environments or in virtual avatar software packages. 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 [1]. One of the challenges in concatenative audiovisual 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. Traditionally, for data-based 2D (audio-)visual speech synthesis (e.g., [1][2]) the information contained in the visual speech database is treated as sequences of static images. To achieve high quality speech synthesis, an accurate analysis of this visual speech data is required. 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.). Such a differentiation is necessary for an optimal visual speech generation. For instance, the overall appearance of the visual speech should be sufficiently smoothed to achieve a natural perception, while on the other hand the movements of the visual articulators should be clearly pronounced to avoid visual under-articulation effects: a strong 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. In this paper we explain why active appearance models are suited for usage with AVTTS and we elaborate on how we achieve audiovisual speech synthesis using a pre-recorded audiovisual speech database and an active appearance model. In addition, we propose some optimization strategies which make use of the specific properties of active appearance models to enhance the quality of the synthesized visual speech.

2. Active Appearance Modeling

2D active appearance models (AAMs) [3] 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} . For some applications, it is convenient that an image is represented by a single set of model parameters, where each parameter determines both shape and texture properties of the image. Therefore, a 'combined' model is calculated from the AAM's shape and texture model, which can be used to transform the image data into a set of 'combined' parameters (and vice-versa). This combined model is determined by the eigenvectors resulting from a PCA analysis on the combination of the eigenshapes P_s and eigentextures P_t . Thus, an image which is projected on the trained AAM can be represented not only by its shape and texture parameters, but also by its corresponding combined parameters. 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 [3].

3. Audiovisual 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 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 [4] 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, in this work we aim to exploit the ability of AAMs to differentiate between shape and texture properties 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 [5] 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 [6][7]. Our final



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

trained AAM was build to retain 97% of the total variation contained in the set of training images and consisted of 8 eigenvectors that represent the shape model and 134 eigenvectors that represent the texture model. The combined model was also calculated, consisting of 94 combined parameters. In addition, we calculated the delta-shape, delta-texture and delta-combined parameters in order to get an estimate for the variation of the parameter values around each video frame. The trained AAM was applied to transform the whole visual database into sequences of shape and texture parameters and into combined parameter trajectories. 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. In addition to the video analysis, the auditory speech in the database was analyzed into acoustic properties like MFCC coefficients, pitch and energy values.

3.3. Segment Selection

The basic concept of our audiovisual speech synthesis strategy has already been described in [1]. The system selects audiovisual segments from an 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 [8]: to synthesize a target sentence, the system searches in the database for audiovisual segments matching the target phoneme sequence. An optimal set of these segments is calculated by means of target costs, which indicate how good a candidate segment matches the target speech, and by join costs, which express how well two consecutive candidate segments can be concatenated without creating join artifacts or abrupt transitions. Since the system searches for appropriate audiovisual segments, both auditory and visual properties are used to calculate the total selection cost of a certain segment. The auditory target cost C_{tar}^{aud} is determined by a set of symbolic binary costs, concerning phonetic and lexical aspects of the candidate and the target segment (e.g., part-of-speech and lexical stress). In addition, a visual target cost is calculated which takes the visual co-articulation effect into account. Due to this co-articulation, the visual appearance of a certain phoneme is greatly dependent on its surrounding phonemes. To calculate this visual target cost, a difference matrix is constructed which expresses the similarity between the visual representation of every two different phonemes present in the database. It is important that this matrix is calculated for the particular database used for synthesis, since the co-articulation effect can behave differently for each speaker. For every different phoneme, all its instances in the database are gathered. For each instance, the AAM combined parameters of the video frame located at the middle of the phoneme are sampled. From these values, the means M_{ij} and variances S_{ij} are calculated, where index i corresponds to the different phonemes and index j corresponds to the different combined

parameters. For a certain phoneme i , the sum of all the variances of the different model parameters $\sum_j S_{ij}$ expresses how much the visual appearance of that phoneme is affected by the visual co-articulation. Two phonemes can be considered similar in terms of visual representation if their mean representations are alike and, in addition, if these mean representations are sufficiently reliable (i.e. if small summed variations were measured for these phonemes). Therefore, two matrices are calculated, which express for each two phonemes the difference between their mean representations and the sum of the variances of their visual representation, respectively:

$$\begin{aligned} D_{pq}^M &= \sqrt{\sum_j (M_{pj} - M_{qj})^2} \\ D_{pq}^S &= \sum_j S_{pj} + \sum_j S_{qj} \end{aligned} \quad (2)$$

Scaling both matrices between zero and one gives $D_{pq}^{M'}$ and $D_{pq}^{S'}$, after which the final difference matrix can be calculated:

$$D_{pq} = 2 \times D_{pq}^{M'} + D_{pq}^{S'} \quad (3)$$

Matrix D_{pq} can be applied to calculate the visual target cost of a certain unit u , matching the target phonetic sequence t : the three phonemes located before ($u+n$) and after ($u-n$) the unit u in the database (i.e., the unit's phonetic context) are compared to the target phonetic context:

$$C_{tar}^{vis} = \sum_{n=1}^3 (D(t-n, u-n)) + \sum_{n=1}^3 D(t+n, u+n) \quad (4)$$

Eq.4 can be further optimized by adding a triangular weighting to the sum. In addition to the target costs, a unit's total selection cost is also determined by auditory and visual join costs. To measure the auditory join cost C_{join}^{aud} , the Euclidean distance between the MFCC coefficients of the audio signals at the join position is calculated. To ensure a smooth visual concatenation, a visual join cost is calculated using Eq.5:

$$C_{join}^{vis} = c_{comb} + c_{\Delta comb} + c_{shape} \quad (5)$$

with c_{comb} the Euclidean distance between the AAM combined parameters of the video frames at the join position, $c_{\Delta comb}$ the Euclidean distance between the AAM delta-combined parameters and c_{shape} the Euclidean distance between the AAM shape parameters. The total selection cost can be calculated by the weighted sum of all sub-costs:

$$C_{total} = w_1 C_{tar}^{aud} + w_2 C_{tar}^{vis} + w_3 C_{join}^{aud} + w_4 C_{join}^{vis} \quad (6)$$

For our system, weights w_1 - w_4 are optimized manually.

3.4. Segment Concatenation

When the most appropriate sequence of segments has been selected, the segments have to be concatenated to create a continuous speech signal. The audio tracks are joined using a pitch-synchronous cross-fade [9]. 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.7 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.7, 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 $1 \leq k \leq m+n-1$: $B_k^j =$

$$\begin{cases} B_k^1 & 1 \leq k < m-S \\ \frac{m-k}{S+1} B_k^1 + \frac{(S+1)-(m-k)}{S+1} B_m^j & m-S \leq k < m \\ \frac{(S+1)-(k-m)}{S+1} B_m^j + \frac{k-m}{S+1} B_{k-m+1}^2 & m < k \leq m+S \\ B_{k-m+1}^2 & m+S < k \leq m+n-1 \end{cases} \quad (7)$$

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. Improving The Synthesis

4.1. Parameter Classification

The AAM-based representation of the database allows to independently characterizing the aspects concerning the shape-information and the aspects concerning the texture-information. However, new possibilities to enhance the synthesis quality emerge when a separate description of different shape- or texture-aspects is feasible. For instance, speech-related changes in shape/texture (e.g., lip movements, teeth visibility, etc.) should be treated apart from the other variations present in the database (e.g., different head orientations among the sentences, illumination changes, etc.). Here we propose a technique to classify each shape/texture parameter in terms of its correlation with the speech. In section 2 it was explained that the model parameters are each linked to an eigenvector, resulting from PCA calculations on the shapes and the textures contained in the training set. 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. To measure the correlation between the parameters and the speech, 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 will 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 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:

$$D_{ij}^{var} = (S_{ij}^{rand} - S_{ij}) / S_{ij}^{rand} \quad (8)$$

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. 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 the mean and the variation of an original trajectory 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 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 should be much alike. In the remainder of this section it will be explained how measures D_j^{var} and D_j^{syn} can be applied to improve the visual speech synthesis.

4.2. Database Normalization

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 in illumination 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. The parameter classification described in section 4.1



Figure 2: *Reconstruction of a database frame using the original model parameters (left) and the normalized model parameters (right)*

can be used to reduce these 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). 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. A subjective evaluation of the effect of the database normalization on the perceived signal quality is given in [6]. It has been shown that the proposed normalization strategy significantly enhances the perceived naturalness of the synthetic speech. An example of an AAM reconstructed frame before and after normalizing the model parameters is shown in Fig.2. Note that the proposed database normalization technique can also be applied to the AAM combined parameters. The normalized versions of these combined parameter trajectories can be used for a more accurate calculation of the visual target- and jointcosts (section 3.3), since in this case only pure speech-related information will be considered to calculate a segment's visual selection cost.

4.3. Differential Smoothing

As was explained in section 3.4, the video concatenations are optimized by smoothing the trajectories at the join positions. This is achieved by overlapping the last frame from the first video segment with the first frame of the second video segment, together with an interpolation of the trajectories around the join position (see Eq.7). A major benefit of the AAM-based synthesis approach is that the amount of smoothing (defined by parameter S in Eq.7) can be diversified between the shape and the texture trajectories: the texture aspects are tweaked more thoroughly than the shape aspects in order to smooth the signal without affecting the articulation strength. To further improve the synthesis, the trajectory smoothing can also be diversified among the shape/texture parameters themselves. Note that the more a parameter is correlated with the target speech, the easier its tweaking will result in unnatural effects. Therefore, the shape parameters as well as the texture parameters were split up into two groups according to their correlation with the speech. While synthesizing a sentence, the parameters of the most speech-correlated group were tweaked using a smaller value of S in comparison with the parameters of the

other group. To determine these two groups, measurements D_j^{var} and D_j^{syn} (see section 4.1) were used. In a way similar to the normalization strategy described in section 4.2, the most speech-correlated parameter group is determined by those parameters which are selected twice as one of the 30% best matching parameters, together with the parameters which are only selected once and which represent more than 1% model variation. In addition, a third measure was added for a more accurate analysis of the texture parameters. We first calculated for each frame the amount of visible teeth and the amount of visible mouth-hole (i.e. the black area inside an open mouth) based on the image’s color histogram. Afterwards, the correlation coefficient between these values and each texture parameter was calculated. For both measures, the 5 most correlated parameters were implicitly added to the speech-correlated parameter group. In addition to the fixed diversification of the smoothing strength among the AAM parameters, each concatenation can be further optimized by tweaking the value of S (for a certain parameter) based on the properties of the phoneme present at the join position. For some phonemes, their visual representation will be more modified by visual co-articulation effects than for other phonemes. As was suggested in [10], the phonemes present in the database can be classified as ‘invisible’, ‘protected’ and ‘normal’. Invisible phonemes (e.g., /t/) are phonemes of which the visual representation is greatly dependant on the visual context of the phoneme instance. This effect can be so pronounced that they are often hardly noticed in the visual speech track. On the other hand, the visual representation of protected phonemes (e.g., /f/) is well-defined and will always be clearly present in the visual speech mode. Since the variability of a phoneme’s visual representation can be speaker dependant, the best strategy is to define these three groups for the particular database used for synthesis. Recall from section 3.3 that the values $\sum_j S_{i,j}$ express the amount of variation of the visual representation of phoneme i . Based on these values, the phonemes of our database were classified as being ‘normal’, ‘protected’ or ‘invisible’. This classification can be used to improve the video concatenation smoothing: when the phoneme at the concatenation point has been classified as ‘invisible’, a heavier smoothing than for ‘normal’ phonemes is applied to avoid an over-articulated visual speech signal and to enhance the signal smoothness of the video track. In contrast, concatenations of a ‘protected’ phoneme are smoothed less profoundly to avoid under-articulation effects. Table 1 summarizes the differential smoothing strategy which is employed in our AVTTS system.

Table 1: *Differential concatenation smoothing*

Type	Speech Correlated	Phoneme Type	S
Shape	High	Protected	1
		Normal	1
		Invisible	3
	Low	Protected	2
		Normal	3
		Invisible	5
Texture	High	Protected	1
		Normal	3
		Invisible	5
	Low	Protected	3
		Normal	5
		Invisible	7

4.4. Spectral Limiting

When the video frames of an audiovisual speech signal are represented by their AAM parameters, the consecutive values of a single parameter throughout a sentence can be seen as a data signal, sampled at the video sampling rate, which contains some portion of the information contained in the visual speech track. The spectral properties of this information can be analyzed by calculating the FFT of the parameter trajectory. We were able to measure that typically, a parameter’s trajectory-spectrum measured in a synthesized sentence contains more energy at higher frequencies in comparison with corresponding trajectories measured in original sentences. This can be explained by the presence of video concatenations: some joins will create changes in shape/texture which are too abrupt and too fast in comparison with natural speech. Note that an accurate concatenation smoothing strategy will not always be sufficient to prevent such unnatural variations, since not all of these rapid changes are caused by concatenation artifacts. Some unnaturally fast variations are due to the fact that the system has selected certain units for consecutive speech targets which are (in their whole) visibly too distant to be put after each other in the output speech track. From our earlier experiments described in [1] and [11], we learned that such rapid variations can cause an over-articulated perception of the synthetic speech. The AAM-based synthesis approach offers the possibility to tackle this problem by tweaking the spectrum of the parameter trajectories: applying a well-designed low-pass filter on the synthesized trajectories will reduce the over-articulation and can enhance the perceived quality of the synthetic visual speech signal. Indeed, such filtering will limit the higher spectral components of the synthesized trajectories which cause the unnatural rapid variations. Note, however, that also the sub-trajectories representing ‘good’ parts of the synthesized speech will be tweaked by the filtering. Since the quality of these parts of the speech should not be affected by the proposed optimization strategy, assessing the effect of the filtering on such ‘good’ trajectories is necessary. Therefore, a perception test has been conducted, for which several original sentences from the speech database were regenerated from their respective AAM trajectories. The participants were shown sample pairs containing different versions of the same utterance, where for one of the two samples the original trajectories were first low-pass filtered before the inverse AAM projection. The position of the filtered sample in each pair was randomized and unknown to the participants. To design the low-pass filters, we first gathered the spectral information of every model parameter of 100 random database movies. The mean of these values gives an estimate for the common spectrum of each parameter trajectory calculated from natural speech. Using these mean spectra, different stopbands were defined for each parameter, preserving 90, 80, 70, 60 and 50 percent of the original spectral information, respectively. For each of these stopbands, a filter has been designed, resulting in 5 different filters for each parameter. For the listening experiment, different combinations of shape and texture filters were applied to the original trajectories. 7 people participated in the test, each comparing 28 sample pairs. They were asked for their preference for one of the two samples in terms of naturalness of the visual speech. Since they were also allowed to answer ‘no difference’, an acceptable filter would result in a high percentage of answers ‘no difference’ or ‘preferred the filtered sample’. A selection of the most important results of the test is given in table 2. To determine which filter to use for a particular parameter while synthesizing visual speech, an approach similar to the differential concate-

Table 2: Perception test on the filtering of original trajectories

shape filter	texture filter	% OK
70	100	86
100	60	93
80	60	64
70	70	79
70	60	36

nation smoothing (section 4.3) can be used: different filter values are applied for the shape and for the texture parameters. In addition, the shape and texture parameters are each split up in two groups based on their correlation with the speech. For the highly speech-correlated parameters, conservative filters are applied (e.g., shape 90% and texture 80%) to avoid over-smoothed visual speech. For the other parameters, a more thorough filtering is applied to enhance the naturalness of the visual output speech (e.g., shape 70% and texture 70%).

5. Discussion

In data-driven AVTTS, one of the major challenges is the construction of a synthetic visual speech mode that appears to be as fluent and as smooth as natural speech. The visual synthesis requires a balanced concatenation smoothing since visual under-articulation should be avoided: the articulation strength of the visual speech should be as equally pronounced as the articulations present in the accompanying (synthetic) auditory speech mode. In this paper we have explained why AAMs are suited to tackle this problem by transforming the information contained in the visual speech database into two sets of parameter trajectories, describing the shape and texture information, respectively. For synthesis purposes, the AVTTS selects an appropriate set of audiovisual segments from the database, using auditory and visual selection costs. In this paper we proposed a visual target and a visual join cost that make use of the AAM parameters to assess a candidate segment's suitability. Once the AVTTS system's unit selection algorithms have selected an appropriate set of video segments, sub-trajectories can be extracted from the database, which are overlapped and interpolated to achieve the video concatenation. 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, several strategies to enhance the quality of the synthetic visual speech were proposed. We described an automatic parameter classification based on a parameter's correlation with the speech. This allows normalizing the visual database by removing database variations which are not due to speech movements. Furthermore, a differential smoothing technique was suggested to further smooth the visual speech without affecting the articulation strength. Manual optimization of the different smoothing strengths showed that an optimal result is indeed achieved when different smoothing strengths are applied. Finally, we proposed an optimization technique where the synthesized parameter trajectories are low-pass filtered to reduce rapid variations that are not found in natural speech. Informal testing showed that for many synthesized sentences, the effect of this filtering is hardly noticed. However, for sentences which do

contain some irregular rapid variations, the filtering does effectively improve the perceived naturalness. Future enhancements to the system should include a text-dependent synthesis of the upper part of the face and the addition of artificial head movements, since it has been shown that an accurate visual prosody enhances the naturalness and intelligibility of the audiovisual speech [12][13]. Examples of some synthesized sentences using the AAM-based audiovisual synthesis approach 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. The authors would like to thank the participants of the listening test for their time.

7. References

- [1] 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
- [2] Cosatto, E and Graf, H.P., "Photo-realistic talking-heads from image samples", IEEE Transactions on multimedia, Volume 2 152–163, 2000
- [3] 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
- [4] 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
- [5] Theobald, B.-J., Fagel, S., Bailly, G. and Elisei, F., "LIPS2008: Visual speech synthesis challenge", Interspeech '08, 1875–1878, 2008
- [6] Mattheyses, W., Latacz, L. and Verhelst, W., "Active Appearance Models for Photorealistic Visual Speech Synthesis", Interspeech '10, 2010
- [7] Stegmann, M.B., Ersboll, B.K., Larsen, R., "FAME - A Flexible Appearance Modelling Environment", IEEE Transactions on Medical Imaging, Volume 22(10), 1319–1331, 2003
- [8] Hunt, A. and Black, A., "Unit selection in a concatenative speech synthesis system using a large speech database", International Conference on Acoustics, Speech and Signal Processing, 373–376, 1996
- [9] Mattheyses, W., Latacz, L., Kong, Y.O. and Verhelst, W., "A Flemish Voice for the Nextens Text-To-Speech System", Fifth Slovenian and First International Language Technologies Conference, 2006
- [10] Woods, J.C., "Lip-Reading: A Guide for Beginners (2nd edn)", Royal National Institute for Deaf people: London, 1994
- [11] Mattheyses, W., Latacz, L. and Verhelst, W., "Multimodal Coherency Issues in Designing and Optimizing Audiovisual Speech Synthesis Techniques", International Conference on Auditory-Visual Speech Processing, Norwich, UK, 2009
- [12] Swerts, M. and Kraemer, E., "The importance of different facial areas for signaling visual prominence", Interspeech '06, 2006
- [13] Munhall, K., Jones, J., Callan, D., Kuratate, T. and Vatikiotis-Bateson, E., "Visual prosody and speech intelligibility", Psychological Science, Volume 15, 133–137, 2004