

# Multimodal Unit Selection for 2D Audiovisual Text-to-speech Synthesis

Wesley Mattheyses, Lukas Latacz, Werner Verhelst and Hichem Sahli

Vrije Universiteit Brussel, Dept. ETRO, Pleinlaan 2, B-1050 Brussels, Belgium

**Abstract.** Audiovisual text-to-speech systems convert a written text into an audiovisual speech signal. Lately much interest goes out to data-driven 2D photorealistic synthesis, where the system uses a database of pre-recorded auditory and visual speech data to construct the target output signal. In this paper we propose a synthesis technique that creates both the target auditory and the target visual speech by using a same audiovisual database. To achieve this, the well-known unit selection synthesis technique is extended to work with multimodal segments containing original combinations of audio and video. This strategy results in a multimodal output signal that displays a high level of audiovisual correlation, which is crucial to achieve a natural perception of the synthetic speech signal.

## 1 Introduction

### 1.1 Text-to-speech Synthesis

A classical text-to-speech (TTS) system is an application that converts a written text into an auditory speech signal. In general, the TTS synthesis procedure can be split-up in two main parts. In a first stage the target text is analyzed by a linguistic front-end which converts it into a sequence of phonetic tokens and the accompanying prosodic information like timing, pitch and stress parameters. Then, in a second step, this information is used by the synthesis module of the TTS system to construct the actual physical waveform. In the early years, model based synthesis was the common technique to create the target speech. This means that the properties of the output waveform are calculated by using pre-defined rules based on measurements on natural speech. For instance, formant-based synthesizers create the synthetic speech by designing a time-varying spectral envelope that mimics the formants found in natural speech. Although these model-based synthesizers are able to produce an intelligible speech signal, their output signals lack a natural timbre to successfully mimic human speech. This led to the development of a different synthesis methodology: data driven synthesizers. These systems construct the target speech by selecting and concatenating appropriate segments from a database with natural pre-recorded speech. If the system can select a good set of segments, the output speech will be perceived as (more or less) natural and it will display a realistic timbre. Currently this data-driven technique is the most common strategy used by high-end TTS systems, where

the segments are selected from a large database containing continuous natural speech signals [12].

## 1.2 Audiovisual Text-to-speech Synthesis

In human to human speech communication, not only the audio but also the visual mode of speech is important. Accordingly, when thinking of a program that converts a written text into a speech signal, ideally this system should create together with the audio a synthetic visual track containing a person that speaks the target text. Such systems are referred to as audiovisual TTS systems. To construct this visual speech signal, the same two major approaches found in classical auditory TTS synthesis exist: model-based and data-based synthesis [1]. Model-based visual speech synthesizers create the visual signal by rendering a 3D model of a human head. To simulate the articulator movements, pre-defined rules are used to alter the polygons of the model in accordance with the target phonetic sequence. Unfortunately, 3D visual speech synthesis systems are unable to produce a completely photorealistic visual speech signal, even when sophisticated models and texture-mapping techniques are used. Similar to the evolution in auditory TTS systems, in recent years more and more interest goes out to data-driven approaches to create a synthetic visual speech signal that is - in the most ideal case - indistinguishable from a natural speech signal. Data-driven audiovisual TTS systems construct the target photorealistic video signal using visual speech data selected from a database containing recordings of natural visual speech. The major downside of data-driven synthesis, both in the audio and in the visual domain, is the fact that the freedom of output generation is limited by the nature and the amount of the pre-recorded data in the database. For instance, the large majority of 2D photorealistic visual speech synthesis systems will only produce a frontal image of the talking head, since their databases consist of frontal recordings only. This means that the system can not be used in, for example, 3D scenery creation in an animated movie. Nevertheless, a 2D frontal synthesis can be applied in numerous practical cases due to its similarity with regular 2D television and video. Research has shown that humans tend to better comprehend a speech signal if they can actually see the talking person's face and mouth movements [18]. Furthermore, people feel more positive and confident if they can see the person that is talking to them. This is an important issue when we think about creating synthetic speech in the scope of machine-user communication. When a TTS system is used to make a computer system pronounce a certain text toward a user, the addition of a visual signal displaying a person speaking this text will indeed increase both the intelligibility and the naturalness of the communication. 2D audiovisual TTS systems are also very useful for educational applications. For instance, small children need a visual stimulus on top of the auditory speech even more than adults do, as it will make them feel more connected with the machine and helps in drawing their attention. Other possible applications can be found in the infotainment sector, where these photorealistic speech synthesizers can be used to create a synthetic news anchor or a virtual

reporter which can, for instance, be employed to create up-to-date audiovisual reports to broadcast via the Internet.

In the remainder of this paper we will focus on data-driven 2D audiovisual TTS synthesis. In the next section we give a general description of this data-driven approach, together with a short overview of the previous work found in the literature. Next, in section 3 we introduce our technique for tackling the synthesis question and we describe our audiovisual text-to-speech system. Our results are discussed in section 4 and section 5 describes how this research can be extended in the future.

## **2 2D Photorealistic Audiovisual Speech Synthesis**

### **2.1 Database Preparation**

Audiovisual data-driven speech synthesis is performed by gathering and combining data from a database. In a first step, an offline recording of the audiovisual database is needed. Note that from the recordings of one single speaker, only one synthetic speaker can be created. This implies that for every virtual speaker we want to create, a new database has to be recorded. In addition, the positioning of the camera determines the possible views of the synthetic head that can be created during synthesis. Another point that needs to be considered is the fact that every head movement of the recorded speaker causes his/her facial parts like the nose, the eyes and the lips to move from their location (when seen from the fixed camera position). So, if we record data including head movements, later on processing will be needed to cope with these displacements. In general, it is not necessary that the audio data is recorded together with the video database, even more: it is not obliged that the same speaker is used. Nevertheless, since the audio and the video mode of an audiovisual speech signal show a great deal of correlation, recording both modes together can have a lot of benefits as will be explained in more detail later. After recording, the database must be analyzed to construct meta-data that can be applied during synthesis. Since this is still an offline step, much effort should be spent on an accurate examination of the speech data because the quality of the synthesis will be for a great deal determined by the nature and the quality of this meta-data. First of all, the speech must be phonetically annotated: the audio signal is segmented in series of consecutive phonemes and the visual signal is segmented in consecutive visemes. In addition, extra properties in both modes are annotated to ensure that the most appropriate segments can be selected during synthesis. Examples of such properties are given in section 3.3.

### **2.2 Speech Synthesis**

To create a new audiovisual speech signal, the synthesizer must select and apply the appropriate data from the database. To create the synthetic audio track, concatenative auditory speech synthesis is the most commonly applied technique. A

general description of this strategy is given in section 3.3 and can be found for instance in [12]. In order to create the synthetic video track, the system has to cope with several requirements. First of all, the synthetic mouth and face movements have to represent the correct phonetic sequence. Note that there is no one-on-one mapping between phonemes and their visual counterpart (visemes): different phonemes can be represented by the same viseme (so-called viseme-classes [2]). On the other hand, due to a strong visual co-articulation effect, several possible visual representations for a same phoneme exist. A second requirement is that the synthetic visual articulators (e.g.: lips, tongue, jaw) should move in a natural manner. Finally, the system must assure that there is a good coherence between the output audio and video mode. In the following section we will briefly describe some techniques that are mentioned in the literature for tackling this synthesis question.

### 2.3 Previous Work

A first important remark that should be made when we inspect the literature on 2D photorealistic speech synthesis is that most of these systems synthesize the audio and the video mode separately from each other. They first acquire the target audio from an external auditory text-to-speech system or from a recording of natural speech and then, afterwards, this audio track and its phonetic transcript are used as input to create the visual mode of the synthetic speech. A second observation is that the systems found in the literature only focus on creating the appropriate mouth movements, after which they complete the synthesis by merging this mouth together with a background face. In the remainder of this paper, although sometimes not explicitly mentioned, we discuss techniques used to synthesize only the mouth-area of the visual speech signal.

In an early system designed by *Bregler et al.* [3], the visual database is first segmented in triphones using the phonetic annotation of the audio track. The system creates a series of output frames by selecting the most appropriate triphones from the database based on two criteria. The first one expresses how well the phonemes of the triphone chunk match the target phonemes: two phonemes from a same viseme-class contribute zero penalty. The second criterion expresses how closely the mouth contours in the boundary frames of the triphone chunk match those in adjacent output frames. Other systems described by *Ezzat et al.* [7] and *Goyal et al.* [10] are based on the idea that the relation between phonemes and visemes can be simplified as a many-to-one relation. First they create a database of still images, one for each viseme-class. For each phoneme in the output audio, its representative still image is added to the output video track. To accomplish a smooth transition between these keyframes, image warping is used to create the appropriate intermediate frames. Much research on 2D speech synthesis was conducted by *Cosatto et al.* [4][5][6]. In the first versions of their system, a map of different mouth occurrences is defined. The different entries of this map are determined by visual properties like the width and the height of the mouth-opening. For each entry, several frames are pre-selected from the database. To synthesize a new visual speech sequence, a trajectory through

the map can be calculated by first training the system with some sample speech data. Then these trajectories are sampled, where the system selects from the target map entries those frames that are most suitable for concatenation. Over the years, their synthesis method evolved more and more towards a real unit selection synthesis, similar to the unit selection techniques used in auditory text-to-speech synthesis. In their approach, the new video track is constructed by selecting and concatenating segments consisting of a variable amount of original frames. This selection is based on how well the fragment matches the ideal target segment and how good it can be concatenated with the other selected chunks.

*Ezzat et al.* [8] and *Theobald et al.* [19] worked on model-based 2D photorealistic synthesis. Their systems first define a model that represents the frames of the visual speech corpus based on shape parameters (e.g.: optical flows or landmarks) and appearance parameters (e.g.: principal components analysis (PCA) coefficients). Such a model can be an analysis tool, since every new frame can be represented as a combination of these shape and appearance parameters. By using such models, the system can generate new unseen frames as every new set of parameters defines a new image. To create the target visual speech signal, trajectories through the parameter space are calculated in accordance with the target phoneme sequence. Based on these trajectories, the system is then able to create a new series of appropriate video frames.

### 3 The Proposed Audiovisual Speech Synthesis System

#### 3.1 General Approach

By developing a Dutch (Flemish) audiovisual speech synthesizer, we wish to investigate how the naturalness of 2D audiovisual TTS synthesis can be further optimized. As explained in section 2.2, the goal in audiovisual speech synthesis is not only to create a visual speech signal that looks fluent and natural, it is also important to reach a high level of multimodal coherence in the output. Since humans are trained to capture and process inputs from both modes of an audiovisual speech input simultaneously, they are sensitive to unnatural combinations of auditory and visual speech. Consequently, the major drawback of the systems described in section 2.3 resides in the fact that they only produce a video signal. Afterwards this signal is merged with an audio track coming from a completely different source (from a different speaker) in order to create the final multimodal output. Although this new video track can appear very natural and smooth, users tend to observe that the auditory speech they hear actually could not have been produced by the facial animation they see. This is often caused by the fact that the visual synthesizer creates a 'safe' representation of the viseme sequence, based on the most common visual representation(s) of the input phoneme sequence. In practice, however, the output audio speech track does include some more extreme phoneme instances (e.g.: badly pronounced ones), which do need a corresponding visual counterpart in the accompanying video track.

In this study, our main goal is to synthesize the output by concatenating audiovisual chunks, selected from an audiovisual database. This means that from the continuous speech in the database, the system will select an appropriate set of multimodal segments from which both the audio and the video track will be used to construct the output speech. This strategy has the advantage that the final output will consist of original combinations of auditory and visual speech fragments, which will maximize the audiovisual correlation in this synthetic signal. This will lead to a more natural perception of the combination of synthetic auditory and synthetic visual speech and it will obviously minimize quality degradations caused by audiovisual co-articulation effects (e.g.: the McGurk effect [16]). In addition, a careful selection and concatenation of the selected audiovisual segments will result in a new multimodal speech signal that exhibits smoothness and naturalness in both its audio and its video mode.

### 3.2 Database Preparation

We recorded a preliminary small audiovisual speech corpus containing 53 sentences from weather forecasts. It is obvious that this limited amount of data will have a negative influence on the overall synthesis quality. Nevertheless, by synthesizing sentences from this limited domain, significant observations are possible. Also, a valorization of the synthesis techniques for the open domain can be attained by expanding the database. The audiovisual speech was recorded with the video sampled at 25 frames per second and the audio sampled at 44100 Hz. We assured that the asynchrony between both modes is negligible small. After recording, the data was analyzed off-line to create the meta-data for synthesis. For the audio track, we computed energy, pitch and spectral properties, together with pitch mark information. The video track was processed to obtain for each frame a set of landmark points, which indicate the location of the facial parts (see figure 1). Additionally, we subtracted from each frame the mouth region and calculated its PCA coefficients. Finally the frames were further processed to detect the amount of visible teeth and the dark area inside the open mouth.

### 3.3 Segment selection

Our audiovisual synthesis system is designed as an extension of our unit selection auditory TTS system [14], which uses a Viterbi search on cost functions to select the appropriate segments from the database. The total cost ( $C_{total}$ ) of selecting a particular audiovisual segment includes target cost functions ( $C_{target}$ ) that indicate how well this segment matches the target speech, and join cost functions ( $C_{join}$ ) which indicate how well two consecutive segments can be concatenated without the creation of disturbing artifacts. To use with our multimodal unit selection technique, these cost functions are needed for the audio track as well as for the video track, since the selection of a particular audiovisual unit will depend on the properties of both these modes. As primary target cost we used the phonetic correctness of the segment. Note that, contrary to the systems described in section 2.3, no viseme-classes are used since the auditory synthesis

requires an exact phonetic match. Since the co-articulation effect - the fact that the visual properties of a certain phoneme strongly depend on the nature of the surrounding phonemes and visemes - is very pronounced for the visual mode, looking for those segments that have a phonetic context matching as well as possible the target speech is crucial. For this reason, the target cost function is further refined to reward a match in the extended phonetic context (see also [14]). To calculate the join cost between two segments, both auditory ( $CA_{join}$ ) and visual ( $CV_{join}$ ) properties are used. For the audio mode, we measure the difference in energy, pitch and mel-scale cepstra. For the visual domain we define an essential cost function that is calculated after aligning the two segments, by measuring the differences between the landmark positions in frames at the border of selected neighboring segments. By using this cost we ensure smooth concatenations in the video mode, since it rewards the selection of mouth instances that are similar in shape. Furthermore, other visual cost functions are needed to select mouths with similar appearances in order to avoid the creation of artifacts at the join positions. This is achieved by comparing properties like the amount of visible teeth and the amount of mouth opening present in the frames. Finally, we implemented a cost function based on the PCA coefficients of the mouth regions, which can be used to measure shape as well as appearance differences.

$$C_{total} = \sum_i w_{target}^i C_{target}^i + \sum_j w_{join}^j C_{join}^j \quad \text{with :}$$

$$\begin{cases} \sum_i w_{target}^i C_{target}^i = w_{phoneme} C_{phoneme} + w_{phoneme-context} C_{phoneme-context} \\ \sum_j w_{join}^j C_{join}^j = \sum_m w^m C_{join}^m + \sum_n w^n C_{join}^n \\ \sum_m w^m C_{join}^m = w_{energy} C_{energy} + w_{pitch} C_{pitch} + w_{mel-scale} C_{mel-scale} \\ \sum_n w^n C_{join}^n = w_{landmarks} C_{landmarks} + w_{teeth} C_{teeth} + w_{mopen} C_{mopen} + w_{PCA} C_{PCA} \end{cases}$$

By adjusting the weights  $w$ , an optimal trade-off between all the different contributions to the total cost can be found. At this point in time, the weights in our system are experimentally optimized, although in a later stage of this research, an automatic training of these parameters might be used to further optimize the selection procedure.

### 3.4 Concatenation and Synthesis

The selected audiovisual segments have to be joined together to create the final output signal. The joining of two fragments that both contain an original combination of audio and video requires two concatenation actions - one for the audio and one for the video track. The two segments that have to be joined are overlapped by an extend that optimizes the concatenation. This join position is first roughly determined by the phonetic transcript of the audio track: former research on auditory speech synthesis has shown that the most optimal join position is the most stable part of the boundary phonemes of the two segments. Each

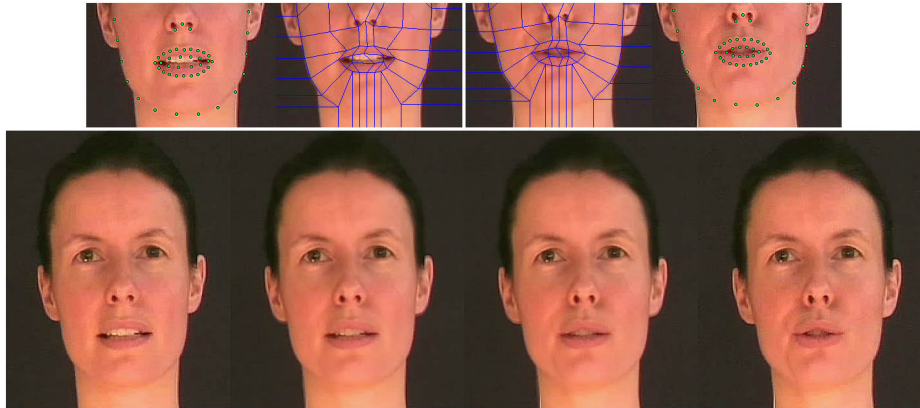
join can be further optimized by fine-tuning this point until the total join cost for this particular join is minimal. In order to successfully transfer the inherent audiovisual coherence from the two audiovisual segments to the joined speech fragment, the location of the join in the video track is kept as close as possible to the location of the join in the audio track (see also further in this section). Joining the two multimodal segments with a certain overlap implies the need for some sort of advanced crossfade technique for both the audio and the video track, as will be explained next.

**Audio Concatenation.** When two voiced speech waveforms are joined, we have to make sure that the resulting signal shows a continuous periodicity. Therefore, we designed a join technique based on pitch mark information that tackles the problem by a pitch-synchronous window/overlap technique. For more details the reader is referred to [15].

**Video Concatenation.** When the video tracks of the two audiovisual segments are played consecutively, we will have to cope with the fact that the transition from the last frame(s) of the first video sequence to the first frame(s) of the second sequence can be too abrupt and unnatural. Therefore, to smooth the visual concatenation, we replace the frames at the end and at the beginning of the first and second video segment, respectively, by a sequence of new intermediate frames. Image morphing is a widely used technique for creating a transformation between two digital images. It consists of a combination of a stepwise image warp and a stepwise cross-dissolve. To perform an image morph, the correspondence between the two input images has to be established by means of pairs of feature primitives. A common approach is to define a mesh as feature primitive for both inputs - so-called mesh warping [20]. A careful definition of these meshes results in a high quality metamorphosis, however, this is not always straightforward and often very time-consuming. Fortunately, when we apply this morphing technique to smooth the visual concatenations in our speech synthesizer, every image given as input to the morph algorithm will be a frame from the speech database and will thus be quite similar to other morph inputs. This means that we only need a strategy to construct the appropriate mesh for a typical frame in the database. To achieve this, we define for each frame a morph-mesh based on the landmarks determined by tracking the facial parts through the database. By using this data as input for the image metamorphosis algorithm, we managed to generate for every concatenation the appropriate new frames that realize the transition of the mouth region from the first video fragment toward the second one. An example of a morph input and the resulting output frames are showed in figure 1.

To construct a full-face output signal, the same technique that can be found in the literature is used (see section 2.3): we first construct the appropriate mouth region signal, which is afterwards merged with a background video showing the other facial parts. Note that some little head movements in the background video have to be allowed, since a completely static head lacks any naturalness. Currently, we only cope with small translations of the background face which

we mimic by carefully aligning the new mouth sequence with the background video.

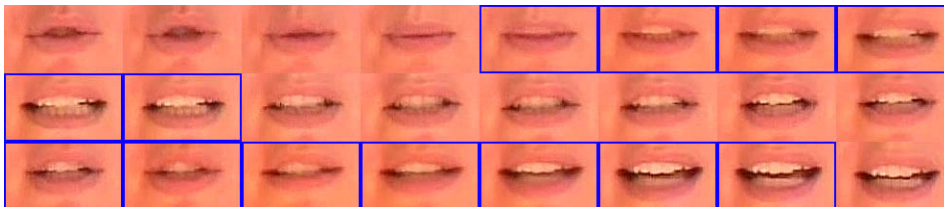


**Fig. 1.** Example of the smoothing technique. The two newly created frames shown in the middle of the lower panel will replace the segments' original boundary frames in order ensure the continuity during the transition from the left frame to the right one. A detail of the landmark data and morph inputs is shown on top.

**Audiovisual Synchronization.** To successfully transfer the original multi-modal coherence from the two selected segments to the concatenated speech, it is important to retain the audiovisual synchronization. In [11], it is concluded that humans are very sensitive to a lead of the audio track in front of the video track in audiovisual speech perception. On the other hand, there is quite a tolerance on the lead of the video signal. In our audiovisual synthesis we exploit this property to cope with the fact that the audio sample rate (44100 Hz) is much higher than the video sample rate (25 Hz). Consequently, the audio component of the selected segments can be joined at exactly the optimal join position (see above), but not so in the video mode whose accuracy is much lower. Therefore, in order to optimize the audiovisual synchrony in the multimodal output signal, at each concatenation, we ensure that the original combinations of auditory and visual speech are desynchronized by maximum half of a frame (20 ms), where a video lead is preferred.

Our system uses the Nextens [13] Dutch linguistic front-end to convert the input text into its phonetic transcript and accompanying prosody information. After concatenation, the sequence of joined segments does not necessarily contain this target prosody. Although we use pitch levels as one of the selection criteria, the concatenated speech will sometimes need extra tweaking to attain the desired output prosody. Therefore, the audio track is processed by a PSOLA algorithm [17] to alter the pitch and the timing of the speech waveform. In order to do so,

a warping path that defines how the timing of the original concatenated signal is mapped on the target timing is constructed. This path is then used to also time-scale the video signal. This visual time-scaling is accomplished by removing or duplicating appropriate frames in such a way that the audiovisual asynchrony remains within the above-mentioned constraints.<sup>1</sup>



**Fig. 2.** From left to right, top to bottom the synthesized mouth sequence for the Dutch phoneme sequence ‘...*O p n e r s l A x s t E i...*’ (coming from the sentence ‘*De kans op neerslag stijgt*’) is shown. Edged frames are newly created ones to smooth the concatenations; the other ones are copied directly from the database.

## 4 General Discussion

Our audiovisual text-to-speech system aims to improve current state-of-the-art audiovisual speech synthesis techniques by increasing the multimodal coherence in the output speech. To achieve this, we apply original combinations of sound/video for concatenation. To select these multimodal fragments, the unit selection paradigm is well-suited since it has been shown to be the most suitable current technique for auditory speech synthesis. Moreover, in [3] as well as in [6] this strategy was successfully applied for the synthesis of the video mode. In contrast with [9], we spent much effort in sophisticating the selection and concatenation process of the synthesizer. Unfortunately, our current speech database is too small to systematically compare and evaluate the overall performance of the system. Nevertheless, preliminary synthesized results from within the limited domain of the database show that an experimentally optimized combination of auditory and visual costs does result in the selection of suitable audiovisual fragments. Furthermore, the two modes of these segments are successfully joined by the proposed multimodal concatenation procedure. An even more important conclusion that can be drawn from the obtained results is that the concatenation of original combinations of audio and video does result in a very high audiovisual coherence in the output signal. This is for instance very noticeable at synthesis points where the selected segments are not optimal. When this results in

<sup>1</sup> We also experimented with more advanced visual time-scaling techniques (e.g.: interpolation by image warping algorithms). However, testing showed that this extra computational workload delivers only very little or even zero gain in signal quality.

some irregularity (non-typical output) in the audio track, the same behavior is noticed in the video track (and vice-versa). More importantly, also at regular synthesis instances the resulting high audiovisual coherence improves the perception of the synthetic multimodal speech, since observers truly believe that the person displayed in the visual mode could indeed have been the source of the presented auditory speech signal. Examples of synthesized sentences can be found at [http://www.etro.vub.ac.be/Research/DSSP/Projects/AVTTS/demo\\_AVTTS.htm](http://www.etro.vub.ac.be/Research/DSSP/Projects/AVTTS/demo_AVTTS.htm). Note that, in order to obtain these results, no manual corrections on the analysis nor on the synthesis were performed.

## 5 Future Work

To further enhance the output quality, the audiovisual database will be enlarged. As found in previous studies on classical auditory text-to-speech systems, providing more initial data to the selection algorithms results in the selection of more optimal units, at the expense of a larger data footprint and a higher computing load. Since audiovisual speech recordings require a lot of data, we expect to find an optimum in this trade-off at about two hours of speech recordings<sup>2</sup>. We will experimentally search for this optimum by conducting listening-tests using databases of variable sizes. Future research will also have to point out new techniques to further optimize the segment selection strategy itself. A first possible enhancement is to tweak the influences of the different cost functions on the total selection cost (e.g.: the importance of visual costs over auditory costs). Another option is to introduce a certain amount of audiovisual asynchrony in order to optimize the concatenations of the segments. Indeed, for each selected audiovisual fragment we could vary the audio and the video join positions independently in such a way that the concatenation can be optimized in both modes separately. Further, aside from the selection of the appropriate mouth segments, a more natural output can be achieved by also altering the movements of the other facial parts in accordance with the input text. Hurdles that will have to be taken to successfully achieve this are the definition of the rules to generate a target visual prosody and the search for a strategy to merge all the different synthesized facial parts into one final, realistic representation of the face.

## 6 Acknowledgments

The research described in this paper was partly sponsored by the IWT project SPACE (SBO/040102) and by the IWOIB project EOS. We would also like to thank Barry-John Theobald for his assistance in landmarking the video in the database.

---

<sup>2</sup> Using a limited-domain approach will always increase the synthesis quality, although this might not be necessary at this database size

## References

1. Bailly, G., Brar, M. and Elisei, F. and Odisio, M.: Audiovisual speech synthesis. *International Journal of Speech Technology* (2003) Volume 6 331–346
2. Breen, A.P., Bowers, E. and Welsh, W.: An Investigation into the Generation of Mouth Shapes for a Talking Head. *International Conference on Spoken Language Processing* (1996) Volume 4 2159–2162
3. Bregler, C., Covell, M. and Slaney, M.: Video Rewrite: Driving Visual Speech with Audio. *Association for Computing Machinery’s Special Interest Group on Graphics and Interactive Techniques* (1997) 353–360
4. Cosatto, E and Graf, H.P.: Sample-Based Synthesis of Photo-Realistic Talking Heads. *Computer Animation* (1998) 103–110
5. Cosatto, E and Graf, H.P.: Photo-realistic talking-heads from image samples. *IEEE Transactions on multimedia* (2000) Volume 2 152–163
6. Cosatto, E, Potamianos, G. and Graf, H.P.: Audio-Visual Unit Selection for the Synthesis of Photo-Realistic Talking-Heads. *International Conference on Multimedia and Expo* (2000) 619–622
7. Ezzat, T. and Poggio, T.: Visual Speech Synthesis by Morphing Visemes (MikeTalk). MIT AI Lab, A.I Memo No: 1658 (1999)
8. Ezzat, T., Geiger, G. and Poggio, T.: Trainable videorealistic speech animation. *Association for Computing Machinery’s Special Interest Group on Graphics and Interactive Techniques* (2002) Volume 21 388–398
9. Fagel, S.: Joint Audio-Visual Units Selection - The Javus Speech Synthesizer. *International Conference on Speech and Computer* (2006)
10. Goyal, U.K., Kapoor, A. and Kalra, P.: Text-to-Audio Visual Speech Synthesizer. *Virtual Worlds* (2000) 256–269
11. Grant, K.W. and Greenberg, S.: Speech Intelligibility Derived From Asynchronous Processing of Auditory-Visual Information. *Workshop on Audio-Visual Speech Processing* (2001) 132–137
12. 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* (1996) 373–376
13. Kerkhoff, J. and Marsi, E.: NeXTeNS: a New Open Source Text-to-speech System for Dutch. *13th meeting of Computational Linguistics in the Netherlands* (2002)
14. Latacz, L., Kong, Y. and Verhelst, W.: Unit Selection Synthesis Using Long Non-Uniform Units and Phoneme Identity Matching. *6th ISCA Workshop on Speech Synthesis* (2007) 270–275
15. 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)
16. McGurk, H. and MacDonald, J.: Hearing lips and seeing voices. *Nature* (1976) Volume 264 746–748
17. Moulines, E. and Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* (1990) Volume 9 453–467
18. Pandzic, I., Ostermann J. and Millen D.: Users Evaluation: Synthetic talking faces for interactive services. *The Visual Computer* (1999) Volume 15 2330–2340
19. Theobald, B.J., Bangham, J.A., Matthews, I.A. and Cawley, G.C.: Near-videorealistic synthetic talking faces: implementation and evaluation. *Speech Communication* (2004) Volume 44 127–140
20. Wolberg, G.: *Digital image warping*. IEEE Computer Society Press (1990)