Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2009, Article ID 169819, 12 pages
doi:10.1155/2009/169819

Research Article

On the Importance of Audiovisual Coherence for the Perceived Quality of Synthesized Visual Speech

Wesley Mattheyses, Lukas Latacz, and Werner Verhelst

Department of ETRO-DSSP, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

Correspondence should be addressed to Wesley Mattheyses, wmatthey@etro.vub.ac.be

Received 27 February 2009; Revised 4 June 2009; Accepted 2 July 2009

Recommended by Gérard Bailly

Audiovisual text-to-speech systems convert a written text into an audiovisual speech signal. Typically, the visual mode of the synthetic speech is synthesized separately from the audio, the latter being either natural or synthesized speech. However, the perception of mismatches between these two information streams requires experimental exploration since it could degrade the quality of the output. In order to increase the intermodal coherence in synthetic 2D photorealistic speech, we extended the well-known unit selection audio synthesis technique to work with multimodal segments containing original combinations of audio and video. Subjective experiments confirm that the audiovisual signals created by our multimodal synthesis strategy are indeed perceived as being more synchronous than those of systems in which both modes are not intrinsically coherent. Furthermore, it is shown that the degree of coherence between the auditory mode and the visual mode has an influence on the perceived quality of the synthetic visual speech fragment. In addition, the audio quality was found to have only a minor influence on the perceived visual signal's quality.

1. Introduction

A classical acoustic text-to-speech (TTS) system converts a written text into an auditory speech signal. In human-to-human speech communication, not only the audio but also the visual mode of speech is important. 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 [1]. Furthermore, people feel more positive and confident if they can see the person that is talking to them. This is an important issue when 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. To construct this visual speech signal two major approaches exist: model-based and data-based synthesis [2]. Model-based visual speech synthesizers create the visual signal by rendering a 3D model of a human head. To simulate the articulator movements, predefined rules are used to alter the polygons of the model in accordance with the target phonetic sequence. Similar to the evolution in acoustic TTS systems, data-driven approaches to create the synthetic visual speech have gained increasing interest over the last years. For instance, some model-based systems try to enhance the naturalness of their output signal by determining the properties of the 3D face mesh and its articulator movements by means of statistical modeling on prerecorded audiovisual speech [3]. Another approach consists of an entirely data-driven synthesis where the output signal is constructed by reusing prerecorded speech data contained in a speech database. Our research focuses on this type of data-driven synthesis, which makes it possible to create a photorealistic video signal that is—in the most ideal case—indistinguishable from a natural 2D speech recording. The major disadvantage of data-driven synthesis is the fact that the flexibility of output generation is limited by the nature and the amount of the prerecorded data in the database. Therefore, the majority of

2D photorealistic visual speech synthesis systems will only produce a frontal image of the talking head as their databases consist of frontal recordings only. Nevertheless, a 2D frontal synthesis can already be applied in numerous practical cases due to its similarity to regular 2D television and video.

2. Motivation

2.1. Previous Work. In an early photorealistic 2D visual speech synthesis system designed by Bregler et al. [4], the visual database is segmented in triphones using the phonetic annotation of the audio track. To create new unseen visual speech, the system creates a series of output frames by selecting the most appropriate triphones from the database. Other systems described by Ezzat and Poggio [5] and Goyal et al. [6] 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. More recent systems use techniques similar to the unit selection strategy found in audio TTS systems. A general description of this strategy can be found in [7]. Cosatto and Graf [8], for example, have created a system where the new video track is constructed by using a visual speech database from which units consisting of a variable amount of original frames are selected and concatenated. This selection is based on how well the unit matches the ideal target speech fragment and how good it can be concatenated with the other selected units. Similar approaches can be found, for example, in [9, 10]. Finally, we should also mention the systems developed by Ezzat et al. [11] and Theobald et al. [12], where the visual speech database is projected onto a model space (e.g., shape and appearance parameters [13]) and where the output speech is constructed by selecting and concatenating model parameters instead of actual frames.

2.2. Motivation. An important observation is that almost all 2D photorealistic visual speech synthesis systems described in the literature synthesize the audio and the video modes of the output speech independently from each other. These systems first acquire the target audio from either an external acoustic text-to-speech system or from a recording of natural speech and afterwards this audio track and its linguistic parameters are used as input to create the visual mode of the output speech. After obtaining the two target speech modes, they are synchronized and multiplexed into one final multimodal output signal. Viewers will capture and process the information contained in the auditory and in the visual speech mode simultaneously. Therefore, any asynchrony and/or incoherency between these two information streams is likely to degrade the perceived quality. Avoiding asynchronies between separately obtained audio and video modes is not straightforward since the synchronization of these two tracks will be based on the original segmentation of the auditory and the visual databases. The segmentation metadata describes the location of the different phonemes in the speech database. In practice, the accuracy of such segmentation information can be rather variable. Therefore, it is possible that the synchronized audio and video tracks contain phonemes of which the visual information appears in video frames that are not played simultaneously with the auditory information of the particular phoneme. At this point, it is unclear what the exact impact of these local and time varying desynchronizations will be on the perception of the multimodal speech signal. From earlier research we do know that for uniform (time invariant) audiovisual desynchronizations even a very small lead of the audio signal is noticed by viewers and causes a degradation of the perceived signal quality [14, 15]. Since in natural speech communication between humans such local asynchronies never occur, it is likely that there exists no such thing as a temporal window in which we are insensitive to audiovisual asynchrony. In addition, such an inaccurate alignment of the two separately synthesized speech modes creates artificial combinations of phonemes and visemes, which can cause various audiovisual coarticulation effects, like the McGurk effect [16]. These effects result in an incorrect perception of the speech information, which degrades the intelligibility of the synthetic speech. Furthermore, even when the two synthetic modes are accurately synchronized, audiovisual incoherencies can still occur in the multiplexed output signal. These are caused by the fact that the auditory and the visual information originates from different repetitions of the same text. Even more, in many of the systems described in the literature, this auditory and visual information is produced by different speakers as these systems use different databases for the acoustic and the visual synthesis. Human speech perception is for a great deal based on predictions, by observing natural speech communication listeners acquired a sense of what is to be considered as "normal" speech. Every aspect of synthetic speech that is not conforming to these "normal" speech patterns will be immediately noticed. Consequently, the different conditions (e.g., phonemic context, prosody, speaker, etc.) from which the synthetic acoustic information (phonemes) and visual information (visemes) originate can result in "abnormal" combinations of auditory and visual speech information that are noticed by a viewer. For instance, some visual speech synthesizers create a "safe" representation of the target viseme sequence, based on the most common visual representation(s) of the input phoneme sequence. In practice, however, the output audio speech track can include some less common phones (e.g., heavily coarticulated consonant clusters). These phones do need a corresponding visual counterpart in the accompanying video track to attain coherent output modes. With our 2D photorealistic text-to-speech synthesis system we aim to investigate how we can create a synthetic audiovisual output signal containing the highest possible coherence between its audio and its video modes. Furthermore, our system can be used to assess the impact of local asynchronies and incoherencies on the perception of the synthetic speech.

3. Multimodal Unit Selection Speech Synthesis

A straightforward solution to increase the degree of intermodal coherence in the synthetic output speech is to synthesize the audio and the video jointly by using prerecorded multimodal speech data. Using the unit selection technique [7], we can select and join audiovisual segments from an audiovisual speech database, such that the final output signal will consist of concatenated original combinations of auditory and visual speech. Consequently, mismatches between the output audio and the output video will be avoided and the intermodal coherence in the output signal will reach almost the same level as found in the natural speech contained in the database. A preliminary study on this approach has been conducted by Fagel [17]. Note that the opposite strategy of synthesizing both modes individually creates more possibilities to optimize the audio and the video, since a separate optimal synthesis strategy and/or database can be designed for either mode. In developing our audiovisual TTS system we wanted to investigate whether the reduced flexibility in design and optimization caused by the joint audio/video synthesis can be justified by the benefits of a maximal audiovisual coherence in the synthetic speech.

3.1. Database. We used the database provided for the LIPS2008 visual speech synthesis challenge [18]. This dataset consists of 278 English sentences, containing auditory and visual speech recorded in “newsreader” style. The data was analyzed offline to create the necessary meta data needed for unit selection synthesis. For the audio track, we computed energy, pitch and mel-scale spectral properties, together with pitch mark information [19]. The video track was processed using an active appearance model (AAM) [13] to obtain for each video frame a set of landmark points, which indicate the location of the face and the facial parts (eyes, nose, upper lip, and lower lip). Additionally, we extracted from each frame the mouth region and calculated its PCA coefficients. Finally the frames were further processed using histogram information to detect the amount of visible teeth and the surface of the dark area inside an open mouth.

3.2. Segment Selection. Our audiovisual synthesis system is designed as an extension of our unit selection auditory TTS system, which uses a Viterbi search on cost functions to select the optimal sequence of long nonuniform units from the database [20]. The cost of selecting a particular audiovisual unit includes target cost functions that indicate how well this segment matches the target speech, and join cost functions that indicate how well two consecutive segments can be concatenated without creating 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. Therefore, the cost c of selecting a particular unit sequence $u_1, u_2, \ldots, u_n$ with corresponding targets $t_1, t_2, \ldots, t_n$ is

$$\begin{aligned} & c(u_1, u_2, \ldots, u_n, t_1, t_2, \ldots, t_n) \\ & \quad = \alpha * \sum_{i=1}^{n} \frac{\sum_{j=1}^{k} w_j^{\text{target}} c_j^{\text{target}}(u_i, t_i)}{\sum_{j=1}^{k} w_j^{\text{target}}} \\ & \qquad + \sum_{i=1}^{n-1} \frac{\sum_{j=1}^{l} w_j^{\text{join\_audio}} c_j^{\text{join\_audio}}(u_i, u_{i+1})}{\sum_{j=1}^{l} w_j^{\text{join\_audio}} + \sum_{j=1}^{m} w_j^{\text{join\_video}}} \\ & \qquad + \frac{\sum_{j=1}^{m} w_j^{\text{join\_video}} c_j^{\text{join\_video}}(u_i, u_{i+1})}{\sum_{j=1}^{l} w_j^{\text{join\_audio}} + \sum_{j=1}^{m} w_j^{\text{join\_video}}} \end{aligned} \tag{1}$$

with c_j^{target} being the target costs and $c_j^{\text{join\_audio}}$ and $c_j^{\text{join\_video}}$ being the join costs for the audio and video concatenation, respectively.

As a primary selection criterion, we used the phonemic correctness of the unit. Typically, this phonemic correctness is not required in visual speech synthesis due to the many-to-one nature of the phoneme to viseme mapping relation, but is obviously necessary in auditory and in multimodal synthesis. Since the coarticulation effect is very pronounced for the visual mode (the visual properties of a phoneme strongly depend on the nature of the surrounding phonemes and visemes), looking for those segments that have a phonemic context matching as well as possible the target speech is crucial. For this reason, one of the target costs rewards a match in the extended phonemic context (see also [20]). Several other target costs are defined, each taking into account a symbolic feature obtained from the linguistic processing front end of the synthesizer [21]. By using a purely symbolic description of the target speech, a detailed prosodic analysis in terms of acoustic values such as f_0 and duration is not required. As prosody prediction is not a straightforward task, it often results in “safe” and thus monotonous predictions in many systems. Therefore, we preferred our purely symbolic approach since it results in more expressive and more natural speech. Examples of symbolic features used in the synthesizer are, for instance, part of speech, lexical stress, and the position in the phrase. For a complete list of these features the reader is referred to [21]. For each demiphone of a candidate unit, its features are compared with those of the corresponding demiphone of the target. Each feature defines a target cost of which the value is calculated by counting the number of demiphones of the candidate unit of which the feature value is different from the target feature value. These target costs can thus be used with units of any size in terms of demiphones.

To calculate the join cost between two segments, both auditory and visual properties are used. For the audio mode, we measure the difference in energy and spectrum (the Euclidean distance between the MFCC’s). Pitch levels are also taken into account by calculating the absolute difference in logarithmic f_0 between the two sides of a join. If the phone at the join position is voiceless, this pitch join cost is set to zero. For the visual mode we define an essential join cost function that is calculated after aligning the two segments that are to be joined, by calculating the Euclidean differences between the aligned mouth landmark positions in the frames at both sides of the join. Other visual cost functions are needed to select mouths with similar appearances in order to avoid the creation of artifacts at the join instants. 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 which calculates the Euclidean difference between the PCA coefficients of the mouth regions at both sides of the join, which can be used to measure shape as well as appearance differences.

3.3. Concatenation. The selected audiovisual segments have to be joined together to create the final output signal. Joining two units containing a combination of audio and video requires two concatenation actions: one for the audio and one for the video track. This implies the need for some sort of advanced cross-fade technique for either of the two modes.

3.3.1. Audio Concatenation. Since we have a series of pitch markers for each audio track, we can exploit the benefits of the use of this pitch information. By choosing a pitch marker as the join instant, we can assure that the periodicity of the speech signal will not be disrupted by the concatenation procedure. The actual concatenation is tackled by a pitch-synchronous cross-fade technique. First, a number of pitch periods (typically 5) are selected around the pitch marker at the end of the first segment and around the marker at the beginning of the second segment. Then, the pitch of these two short segments is altered using the PSOLA technique [22], which will result in two signals having exactly the same pitch. The initial pitch value of these resulting signals is chosen equal to the pitch present in the original signal extracted from the first segment. This pitch then varies smoothly along the length of the signals such that the final pitch value becomes equal to the pitch of the signal extracted from the second segment. Finally, these two completely pitch synchronized signals are cross-faded using a hanning function to complete the concatenation. This strategy minimizes the introduction of irregular pitch periods and assures the preservation of the periodicity as much as possible. For more details the reader is referred to [23].

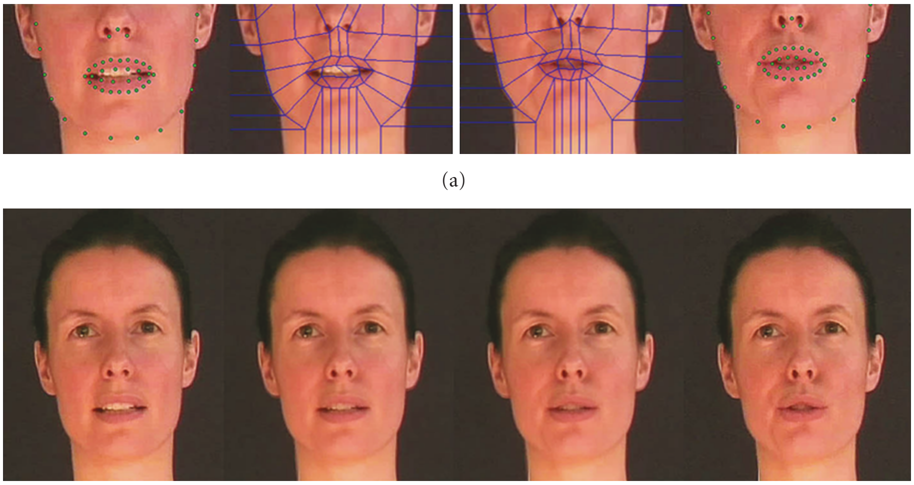
3.3.2. Video Concatenation. When the video tracks of the two audiovisual segments are played consecutively, we will have to cope with the fact that the transition from the last frame(s) of the first video sequence to the first frame(s) of the second sequence can be too abrupt and unnatural. Therefore, to smooth the visual concatenation, we replace the frames at the end and at the beginning of the first and second video segments, respectively, by a sequence of new intermediate frames. Mesh-based image morphing is a widely used technique for creating a transformation between two digital images [24]. A careful definition of the two meshes used as feature primitives for both images results in a high quality metamorphosis. We define for each frame of the database a morph mesh based on the landmarks determined by tracking the facial parts. By using this data as input for the image metamorphosis algorithm, we managed to generate for every concatenation the appropriate new frames (typically 2) that realize the transition of the mouth region from the first video fragment toward the second one (see Figure 1).

To create a full-face output signal, we first construct the appropriate mouth region in accordance with the target speech as described above. Afterwards, this signal is merged with a background video showing the other parts of the face. At this point, we did not yet investigate a strategy to mimic an appropriate visual prosody in the background video. Since it has been shown that there exists some level of synchrony between the movements of the head/eyebrows/eyes and the linguistic/prosodic properties of the speech [25, 26], we should avoid providing the output speech with a random visual prosody. Therefore, we created a background signal displaying a neutral prosody with only very little head movements and one repetitive eye blink. This will prevent the users from being distracted by inappropriate movements, while on the other hand this will be perceived as much more natural than a completely static frame as background (see Figure 2).

3.4. Audiovisual Synchronization. To successfully transfer the original multimodal coherence from the two selected segments to the concatenated speech, it is important to retain the audiovisual synchronization. In [15], 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 optimize the concatenation of the selected audiovisual segments. In order to join two segments, we introduce a certain degree of overlap. For each concatenation, the exact join position is determined by examining the audio tracks and selecting the pair of pitch mark instants that minimizes the auditory join cost for this particular join [27]. Since the sample rate of an audio signal is much higher than the sample rate of a video signal, the join position in the visual mode cannot be determined with the same accuracy. In order to optimize the audiovisual synchrony in the multimodal output signal, for each concatenation the video join position is located as closely as possible to the join position in the audio track. In addition, we ensure that throughout the whole output signal the original combinations of auditory and visual speech are desynchronized by the smallest possible video lead, that is, between zero and one video frame (40 ms for a 25-femtosecond video signal).

4. Experiments

In this section we describe the experiments we conducted in order to assess the impact of the joint audio/video synthesis on the quality of the synthesized speech. Note that the assessment of the quality of audiovisual speech covers different aspects, as there are intelligibility, naturalness,



(a)

(b)

Figure 1: Example of the video concatenation technique using our Dutch audiovisual database. The two newly created frames shown in the middle of (b) will replace the segments' original boundary frames in order to 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 (a). Note that at the end only the mouth area of these frames will be used in an overlay on the background video.

and acceptance ratio measures. In Section 2.2 we explained how audiovisual mismatches could lead to coarticulation issues that decrease the intelligibility of the synthetic speech. Furthermore, we discussed the negative consequences of audiovisual asynchronies and incoherencies on the perceived naturalness of the speech. Our multimodal selection strategy was designed to ensure a high multimodal coherence in the output signal, which on the other hand reduces the flexibility in selection and optimization in comparison to a separate synthesis of both modes. Therefore, it was necessary to evaluate whether this limitation can be justified by the positive effects of such a maximal intermodal synchrony and coherency. Our experiments are designed to find out whether the joint audio/video synthesis indeed results in a minimization of the auditory-visual mismatches and to assess the consequences for the perceived naturalness of the synthetic speech. Thus, the impact of the joint audio/video synthesis will be measured both directly and indirectly. For a direct assessment we should measure to which extend a viewer notices mismatches between the two modes of audiovisual speech synthesized using different strategies. Moreover, we can also indirectly measure the effect of the reduction of audiovisual mismatches as a result of the joint audio/video selection by assessing its impact on the perceived naturalness of the synthesized speech. To do so, we designed a listening test containing two experiments in which we measure the effects of the multimodal unit selection synthesis directly and indirectly, respectively. If the results of the listening test point out that the high degree of coherence between the speech modes synthesized using the joint audio/video selection technique does indeed have a significant positive impact on the perceived quality, the reduced flexibility for selection and optimization is warranted and further optimizations on the joint audio/video synthesis strategy should be investigated.

4.1. Experiment 1

4.1.1. Goal. In a first experiment we measured the detection of audiovisual mismatches between the two modes of synthetic audiovisual speech. These mismatches can be classified as either *synchrony issues* (caused by an inaccurate synchronization of the two signals) or as *incoherency issues* (caused by the different origin of the auditory and the visual information), as was discussed in Section 2.2. Although it is very hard to directly detect such *incoherencies* in continuous speech, it is possible for a viewer to detect certain local auditory-visual *asynchronies* and thus to rate the overall synchrony between two presented speech modes. In this experiment we examined if there is any difference in the reported synchrony for audiovisual sentences synthesized by the joint audio/video selection technique and sentences of which both modes are synthesized separately. For the latter we also examined whether there is a difference when the databases used for the auditory and for the visual synthesis are the same or different.

4.1.2. Method. Audiovisual sentences were displayed to the subjects which were asked to rate the overall level of synchrony between the audio and the video tracks (i.e., to assess whether the viewers did notice some local audiovisual asynchronies). Since some of the subjects were nonspeech experts, we gave the participants the extra advice that synchrony issues are typically noticeable at mouth openings for vowel instances and at plosives. It was stressed that they should rate only the level of time synchrony, and not, for instance, the smoothness or naturalness of the signals. The subjects were asked to use a 5-point MOS scale, with rating 5 meaning "perfect in synchrony" and rating 1 meaning "large asynchronies noticed". There was no time limit and the viewers could play and replay each sample any time they wanted. The mean time the participants spent for both experiment 1 and experiment 2 (see Section 4.2) was about 40 minutes. A short break of a few minutes was provided between the two tests. The video samples were presented on a standard LCD screen, placed at normal “office working” distance from the viewers. The video signals were 532 × 550 pixels large and they were displayed at 100% size. The audio signal was played through high-quality headphones.

4.1.3. Subjects. Eleven subjects participated in this test, seven of which were experienced in speech processing. Six of the subjects were aged between 20–30 years, the other subjects were between 35–57 years of age (mean age = 36). The group of participants consisted of 3 female and 8 male subjects. None of them was native English speakers: 8 of them were Dutch speaking; the other participants were Chinese, Greek and Turkish. We did, however, assure that all participants had good command of the English language.

4.1.4. Synthesis Strategies. Four types of speech samples were used for this test (see Table 1), with each sample containing one average length English sentence. The first group (ORI) contained natural audiovisual speech samples selected from the LIPS2008 database. A second group of samples (MUL) were synthesized using the multimodal selection and concatenation strategy discussed in the previous section of this paper. This means that these samples consisted of concatenated original combinations of audio and video. As explained in Section 3.4, these original combinations are desynchronized by the smallest possible video lead, which should be unnoticeable for the participants. The AVTTS system was provided with the LIPS2008 audiovisual database from which the particular sentence that was to be synthesized was excluded. The selection costs were tweaked to maximize the quality of the visual mode by raising the weights of the visual join costs and by lowering the weights of the auditory join costs. Likewise, sentences can be synthesized using an opposite “best audio quality” tweaking of the costs in favor of the auditory quality. The third group of test samples (SAV) was created by joining the audio track of such a “best audio quality” synthesis together with the video track of the “best video quality” synthesis that was also used for the creation of the (MUL) samples. The audiovisual synchrony was assured by performing a nonuniform time scaling on the audio track using WSOLA [28] in order to align it with the video track. Thus, although they are selected from the same database, the auditory and the visual speech modes of the (SAV) samples are not intrinsically coherent as it is the case for the (MUL) samples. A fourth set (SVO) of sample sentences were created in the same way as the (SAV) set, but with a different system to construct the auditory speech; the audio track of these samples was created by using our auditory text-to-speech system [21] provided with the CMU ARCTIC database [29] of an English female speaker. This database is commonly used in TTS research and its length of 52 minutes allows higher quality audio synthesis than that of the LIPS2008 database. The resulting audio track was then also time-scaled and joined with the video of a “best video quality” synthesis as described before. Note that this audiovisual synthesis strategy is similar to most other audiovisual text-to-speech systems found in the literature, where different systems and databases are used to create the audio and the video tracks of the output signal. All samples, including the files from group (ORI), were (re-)coded using the Xvid codec with the same quality settings, resulting in a homogeneous picture quality among all samples. Note that all files were created fully automatically and no manual correction was involved for any of the synthesis or synchronization steps.

4.1.5. Samples. We synthesized 15 sample sentences with mean word count of 15.8 words using the settings for each of the four groups (ORI, MUL, SAV, and SVO) as described above. Each viewer was shown a subset containing 20 samples: 5 different sentences each synthesized using the four different techniques. While distributing the 15 sample sentences among the participants, each sentence was used as many times as possible. The order in which the samples of a certain sentence were shown was randomized.

4.1.6. Results. Table 2 shows the summary of the test results for each group. In Figure 3, the results of the experiment are shown by means of a box plot.

It is generally accepted that the statistical analysis of MOS ratings should consist of nonparametric tests, since a MOS scale does not exhibits the properties of an equal interval scale. Therefore, we conducted a Wilcoxon test to every pair of test groups, from which the resulting P values are shown in Table 3.

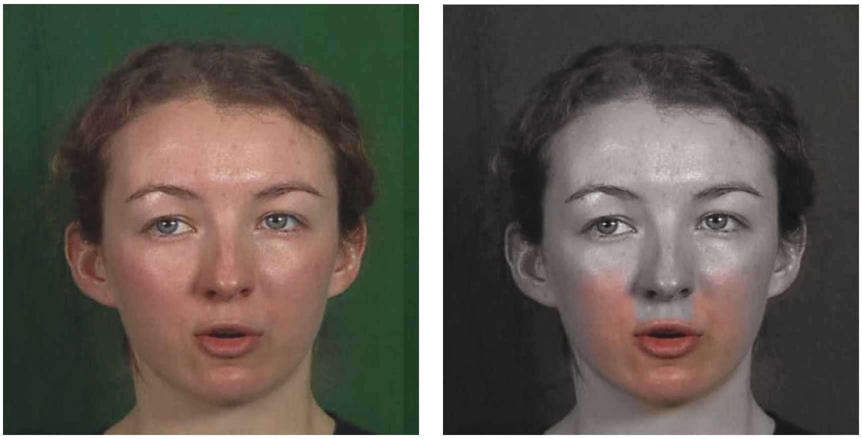
Using a significance threshold level $\alpha = 0.05$, Bonferroni corrected to $\alpha = 0.0083$, we get in Table 4 significant differences between the sample groups.

Further analysis of the test results showed no difference between the overall ratings of the speech experts and the ratings given by the nonspeech experts. The female participants reported slightly better ratings, although a Mann-Whitney test showed that this difference is only significant to the 0.85 level of confidence. Note that this difference is likely to be caused by the limited number of female viewers in comparison to the amount of male participants. A Kruskal-Wallis test showed that the overall ratings differ among the participants with a significance of 0.98; some participants reported in general higher ratings than other participants. Maybe we could have prevented this by introducing some training samples which indicate a “best” and “worse” sample. However, in this test we were mostly interested in the pairwise comparisons among the different synthesis strategies for a single sentence rated by each particular viewer.

4.1.7. Discussion. For each group of samples, an approximation of the actual audiovisual synchrony can be made. For group (ORI), a perfect synchrony exists, since no significant audio or video lag was present in the database recordings. Samples of group (MUL) are made out of concatenated original combinations of audio and video. This implies that most of the time they exhibit the original synchrony as found

Table 1: Different synthesis strategies used in the experiments.

GROUP	ORI	MUL	SAV	SVO	RES
Origin audio	Original LIPS08 audio	"Best video" unit selection on LIPS08 database	"Best audio" unit selection on LIPS08 database	Acoustic unit selection on ARCTIC database	Original LIPS08 audio
Origin video	Original LIPS08 video	"Best video" unit selection on LIPS08 database	"Best video" unit selection on LIPS08 database	"Best video" unit selection on LIPS08 database	"Best video" unit selection on LIPS08 database
Description	Natural AV signal	Concatenated original AV combinations	Separate synthesis on same database	Separate synthesis on different database	Synthesized video and original audio



(a)

(b)

Figure 2: (a) displays an example frame from a synthesized sentence from the (MUL) group. (b) shows the area that is synthesized in accordance with the text (colored) and the background signal containing a neutral visual prosody (grey).

Table 2: Summary of the test results for experiment 1 (a 5-point MOS).

	ORI	SVO	SAV	MUL
Median	5	2	3	3
Mean	4.88	2.26	3.19	3.30

Table 3: *P*-values of a Wilcoxon test on the results of experiment 1.

	ORI	SVO	SAV	MUL
ORI		3.97e-8	3.29e-8	8.81e-8
SVO			4.00e-5	5.52e-5
SAV				0.561
MUL				

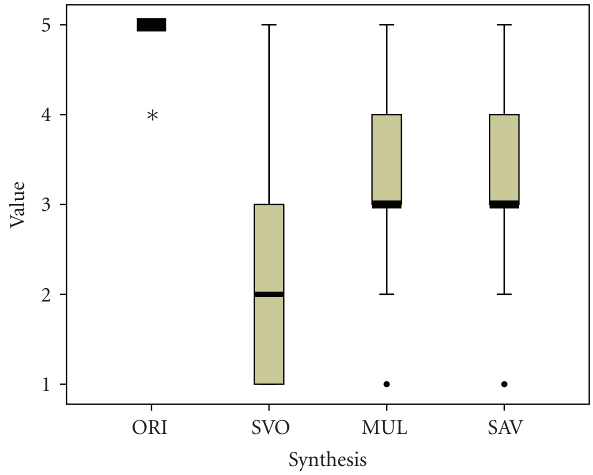


Figure 3: Box plot of the results of experiment 1.

in the database recordings. Only for the frames at the join instants an exact definition of the synchrony is impossible since at these moments the signal consists of an interpolated audio track accompanied by an interpolated video track. For the (SAV) and the (SVO) samples, we tried to align both modes as accurately as possible, as was verified by manual inspection of the signals. The results of the experiment show that the perceived audiovisual synchrony does differ between the groups. There is a significant difference between the ratings for group (ORI) and group (MUL); it seems that it is hard for a viewer to assess only the audiovisual synchrony

Table 4: Significant differences for experiment 1, $\alpha = 0.05$ Bonferroni corrected to $\alpha = 0.0083$.

	ORI	SVO	SAV	MUL
ORI		TRUE	TRUE	TRUE
SVO			TRUE	TRUE
SAV				FALSE
MUL				

without being influenced by the overall smoothness and naturalness of the signals themselves. Also, the perception of the synchrony of the (MUL) samples could be affected by the moderate loss of multimodal coherence at the join instants, where the speech consists of interpolated audio and video signals. Between groups (MUL) and (SAV) no significant difference was found. At this point we should remark that the samples of groups (MUL) and (SAV) are more similar than one would expect. The reason for this is two fold. First of all, our multimodal unit selection strategy requires the selected segments to be phonemically identical with the targets. This is in contrast with video-only synthesizers, where only viseme similarity with the target speech is needed, resulting in more candidate units per viseme. Since the LIPS2008 database is small (around 25 minutes of speech) compared to other unit selection databases (often more than 2 hours of speech), the amount of available candidate units will be quite limited for our synthesis approach. Furthermore, we found that for this database, the best results were obtained when the synthesizer selects the longest units possible (instead of selecting more but smaller candidate units, see [20]). This implies a further decrease in number of candidate units since for many of the long (more than three phonemes) units only one candidate will be present. Analysis of the units selected by "best video quality" and "best audio quality" syntheses of the same sentences showed that sometimes up to 70% of the selected units were the same. Given the fact that these are often long units, we calculated that for the (SAV) samples, although the audio and the video were selected using different settings, on average around 50% of the video frames are accompanied by the original matching audio from the database. For some sentences this number even increased up to 85% of the frames. This implies that we should indeed not expect much difference in test results from these two groups. However, the ratings for the (MUL) and the (SAV) samples prove that the technique used to synchronize the two separately synthesized modes is capable to successfully align the two signals. Otherwise, the ratings for the (SAV) samples would have been worse in comparison to the ratings for the (MUL) samples. A significantly better rating was found for the (MUL) group in comparison to the (SVO) group. This indicates that the participants, although underestimating the synchrony of the (MUL) samples, do notice the difference between the intrinsic natural synchrony of the (MUL) group and the simulated synchrony of the (SVO) samples. As the alignment algorithms were found to successfully synchronize the separately synthesized modes, a possible reason for the worse ratings for the (SVO) group could be that viewers underestimate the synthetic synchrony of the (SVO) group because they are sensitive for the overall lack of intermodal coherence in these samples. Since both tracks are produced by using different databases from different speakers, the resulting signal will sometimes contain a visual speech fragment for which it is hard to believe that it could have been the source of the accompanying auditory speech (in spite of both tracks separately being of acceptable quality). Apparently, these mismatches are perceived by the participants in a similar fashion as local auditory-visual asynchronies and will thus cause a degradation of the perceived multimodal synchrony.

4.2. Experiment 2

4.2.1. Goal. In a second experiment we assessed the effect of the multimodal selection technique on the perceived naturalness of the audiovisual speech. Note that the naturalness of an audiovisual speech signal is determined by the naturalness of the individual audio and video mode. In addition, it is affected by the naturalness of the combination of these two modes. For instance, a video track which exhibits high quality synthetic visual speech (i.e., the movements of the visual articulators are smooth and in accordance with the text as in true natural speech) could be perceived as much less natural when it is played along with a badly matching audio track. The naturalness of the combination of an audio and a video modes can be enhanced by minimizing the intermodal incoherence issues, which we aim to realize with the joint audio/video synthesis approach. In order to measure this effect, we should evaluate the perceived naturalness of different audiovisual speech samples, where the individual qualities of the audio and of the video modes are constant and where a variation in multimodal coherence exists among the samples. However, it is not clear how we could realize such samples in practice. Therefore, we created several groups of audiovisual speech signals, each synthesized using different synthesis strategies. We ensured that for every group, the visual mode was constructed by the concatenation of the same video segments. Thus, the quality of the visual mode is the same for all groups. By comparing the perceived naturalness of these video tracks, played along with different types of auditory speech, the effect of a high degree of multimodal coherence on the perceived naturalness of the visual speech (and thus on the overall audiovisual naturalness) can be measured. Furthermore, these measurements can also be used to evaluate the impact of the quality and the naturalness of the auditory speech on the perceived naturalness of the accompanying video track. It is interesting to know whether this effect is as important as the impact of the level of auditory-visual coherence.

4.2.2. Method and Subjects. In this test the participants were asked to rate the naturalness of the mouth movements displayed in audiovisual speech fragments. A 5-point MOS scale was used, with rating 5 meaning "the mouth variations are as smooth and as correct as natural visual speech" and rating 1 meaning "the movements considerably differ from the expected natural visual speech". The same subjects as in experiment 1 participated in this test.

Table 5: Summary of the test results for experiment 2 (a 5-point MOS scale).

	ORI	SVO	SAV	MUL	RES
Median	5	2	3	3	3
Mean	4.91	2.53	3.11	3.37	3.01

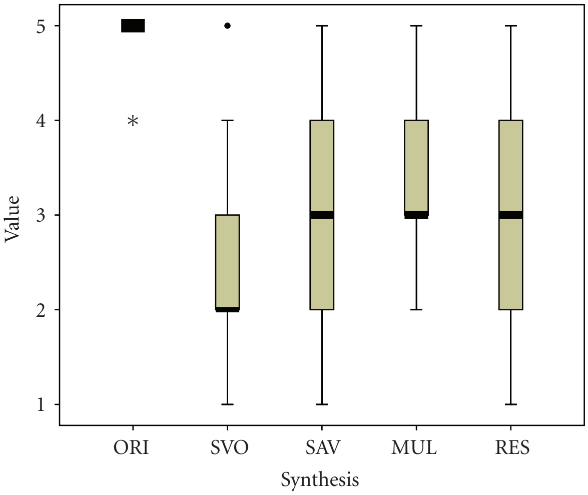


Figure 4: Box plot of the results of experiment 2.

4.2.3. Synthesis Strategies. For this test we used the same groups of samples as we used in the first experiment (ORI, MUL, SAV, and SVO), augmented with a fifth group (RES) containing sentences from which the audio mode is an original recording from the LIPS2008 database. The video mode of these samples was synthesized using the “best video quality” settings and the LIPS2008 speech database from which the particular sentence was excluded. Afterwards, both modes were aligned and joined in the same way as we did for the (SVO) and (SAV) samples. Note that this synthesis method can be seen as a special case of audio-driven visual speech synthesis, where the auditory speech used as input and the video database used for synthesis are recordings from the same speaker.

4.2.4. Samples. We used the same 15 English sentences as were used in the first experiment. Every sentence was additionally synthesized using the (RES) strategy. Each subject was shown a subset containing 20 samples: 4 different sentences, each synthesized using the 5 different strategies from Table 1. While distributing the 15 sample sentences among the participants, each sentence was used as many times as possible. The order in which the samples of a given sentence were shown was randomized.

4.2.5. Results. Table 5 shows the summary of the test results for each group. In Figure 4, the results of the experiment are showed by means of a box plot.

Again, we performed a Wilcoxon test to every pair of test groups, from which the resulting P-values are shown in Table 6.

Table 6: P-values of a Wilcoxon test on the results of experiment 2.

	ORI	SVO	SAV	MUL	RES
ORI		1.77e-8	2.21e-8	2.88e-8	7.13e-9
SVO			0.0118	0.00140	0.0296
SAV				0.210	0.436
MUL					0.0474
RES					

Table 7: Significant differences for experiment 2, $\alpha = 0.05$ Bonferroni corrected to $\alpha = 0.005$.

	ORI	SVO	SAV	MUL	RES
ORI		TRUE	TRUE	TRUE	TRUE
SVO			FALSE	TRUE	FALSE
SAV				FALSE	FALSE
MUL					FALSE
RES					

Using significance threshold level $\alpha = 0.05$, Bonferroni corrected to $\alpha = 0.005$, we get in Table 7 significant differences between the sample groups.

Further analysis of the test results showed that the overall ratings of the nonspeech experts were slightly higher than the ratings given by the speech experts (Mann-Whitney test significance = 0.8). Furthermore, similar to the results of the first experiment, the female participants reported slightly better ratings in comparison to the male subjects (Mann-Whitney test significance = 0.9). As the subjects were the same as for the first experiment, this difference is again likely to be caused by the limited number of female viewers in comparison to the amount of male participants. Inspection of the overall ratings of each participant showed that in this second experiment like in the first one some participants reported in general higher ratings than other participants (Kruskal-Wallis test significance = 0.99). Nevertheless, also for this test we were mostly interested in the pairwise comparisons among the synthesis strategies for a single sentence rated by each particular viewer.

4.2.6. Discussion. For all but the (ORI) samples, the visual mode was synthesized by using the LIPS08 database and the same “best video quality” settings. This implies that any significant difference in perception quality of the visual speech will be caused by the auditory speech played along with the visual mode. By comparing the (MUL) results to the (SVO) results, a clear preference for the (MUL) samples is noticeable ($P = .0014$). Note that the quality of the separate audio mode of the (SVO) samples is at least as high as the quality of the audio mode of the (MUL) samples, since the (SVO) samples are synthesized using only acoustic selection costs. In addition, the ARCTIC database is much larger than the LIPS08 database which results in more candidate units for synthesis. This implies that the perceived naturalness of the visual speech of the (SVO) sentences was degraded by the artificial combinations of audio/video present in these samples. This indicates that the minimization of such intermodal mismatches will have a profound positive influence on the perceived overall naturalness. In contrast, only a little decrease in perceived naturalness is noticed between the (MUL) and the (SAV) samples. As explained earlier, for both groups the audio quality and the degree of audiovisual coherence are probably too similar to cause noticeable perception differences. It also shows that the synchronization of the two separately synthesized modes was realized with appropriate accuracy. Further analysis of the test results shows that there exists a difference between the ratings for the (MUL) samples and the ratings for the (RES) samples. Since the audio track of the (RES) samples contained natural auditory speech, an optimal perception of these video tracks could be expected. However, the results indicate that the viewers gave a higher rating to the samples of the (MUL) group ($P = .047$). Despite the fact that this P-value is above the significance threshold, these ratings show that for a high quality perception of the visual speech mode, a high degree of audiovisual coherence is equally or even more important than the individual quality of the auditory speech. In addition, it is also worth mentioning that by comparing the (SVO) to the (RES) samples, there is an indication ($P = .030$) that the higher quality of the auditory speech does have a positive influence on the perception of visual speech. However, compared to the influence of the multimodal coherence, this effect is only secondary. From the results obtained in this second experiment we can conclude that audiovisual speech synthesis strategies should ensure an optimal auditory-visual coherency in order to attain an output signal that is perceived to be natural. Obviously, the individual quality of the audio and video track is important as well, but the experiments show that the perception of the combination of individually optimized auditory and visual speech modes will be only suboptimal when multimodal coherency issues are present in the output signal.

5. Conclusion

In this paper we have described our strategy to perform audiovisual text-to-speech synthesis. We adopted the unit selection method to work with multimodal units, using audiovisual selection costs. This strategy makes it possible to create multimodal speech signals of which the synthetic audio mode and the synthetic video mode are highly coherent. This differs from most strategies found in the literature, which use completely separated systems, methods and databases to construct the auditory and the visual mode of the output speech.

We conducted two experiments in order to assess the influence of this strong multimodal coherence on the perception of the synthetic visual speech. In a first test we measured the perceived audiovisual synchrony resulting from different synthesis strategies. It showed that viewers tend to underestimate the audiovisual synchrony when the displayed signals are synthetic and distinguishable from natural speech; this may be due to a moderate loss of coherence due to the interpolation mechanisms employed at the audio and video segment joins and/or to unnatural overall variations and prosody in the synthetic signals. On the other hand, the audiovisual signals created by the multimodal selection technique are perceived as more synchronous compared to the signals of which both modes are constructed separately and synchronized on the phoneme level afterwards. Apparently there exists a decrease in perceived synchrony when the test subjects (unconsciously) notice some mismatches between the audio mode and the video mode. Moreover, in the second experiment we indirectly measured the audiovisual coherence by evaluating the perceived naturalness of the visual speech mode of different syntheses. This test showed that a synthetic visual speech fragment is perceived as more natural when there is a strong coherence between the visual speech and the auditory speech playing along. The influence of the individual quality of the accompanying auditory speech on the perceived naturalness of the visual speech seems to be only of secondary order.

From the two experiments we can conclude that a separate synthesis of the audio and the video track, using different techniques and different databases, is likely to cause multimodal incoherencies which cannot be eliminated by an accurate synchronization of the two signals, since they are due to the fact that the two information streams originate from different repetitions of a same utterance by different speakers. Since it has been shown in our experiments that these mismatches reduce the perceived synchrony and naturalness of the synthetic speech, audiovisual speech synthesis strategies should be designed in order to minimize these incoherencies. This is found to be at least as important as the optimization of the individual quality of the auditory and the visual speech. The multimodal selection technique proposed in this paper is able to do so; it maximizes the intermodal coherence at the expense of a decrease in selection and optimization flexibility. The most straightforward solution for this loss could be to extend the audiovisual database used for synthesis. In addition, the individual quality of both modes can be further optimized by improving the joint audio/video selection costs and the multimodal concatenation techniques. Note that any optimization to the synthesis should be designed in such a way that it does not result in a loss of intermodal coherence in the output speech.

From the results obtained we believe it is important to further investigate the importance of the coherence between audio and video modes for the perceived quality and naturalness of audiovisual speech synthesis and other applications such as audiovisual speech recognition and audio-to-audiovisual speech mapping techniques. The experiments conducted show that there is a significant impact of the accompanying auditory speech on the perceived visual speech quality. Thus, it could be an interesting option to also involve different audio speech tracks in challenges as LIPS [18], where the quality of synthesized visual speech among the participating systems is assessed.

While our experiments clearly showed that the choice for coherent auditory and visual segments will improve the perceived naturalness, at this point the exact impact of selecting the audio and the video fragments separately but from a same audiovisual database is still unclear. Future experiments using a larger database with more candidate units will hopefully answer this question. We should also investigate the impact of the different synthesis strategies on the overall quality of the synthetic audiovisual speech. For the second experiment the participants assessed the quality of the visual speech only. If, in contrast, we would ask to rate the combined auditory and visual speech quality, it is likely that the audio-driven synthesized samples would get a better rating since their audio track consists of natural speech. Note, however, that the results described in this paper illustrate that the outcome of such experiments is hard to predict since many intermodal effects can have an influence on the perception of an audiovisual speech signal. Sample syntheses created by the multimodal unit selection technique can be found on our website: http://www.etro.vub.ac.be/Research/DSSP/Projects/avtts/demo_avtts.htm.

Acknowledgments

The research reported on in this paper was supported in part by the Institute for the Promotion of Innovation by Science and Technology in Flanders project SPACE (IWT-SBO/040102): SPeech Algorithms for Clinical and Educational Applications and by a research grant from the Faculty of Engineering Science, Vrije Universiteit Brussel. The authors thank the reviewers for their comments that helped to considerably improve the original manuscript. They would also like to thank the people who participated in the experiments for their time.

References

[1] I. S. Pandzic, J. Ostermann, and D. Millen, "User evaluation: synthetic talking faces for interactive services," *The Visual Computer*, vol. 15, no. 7-8, pp. 330–340, 1999.

[2] G. Bailly, M. Brar, F. Elisei, and M. Odisio, "Audiovisual speech synthesis," *International Journal of Speech Technology*, vol. 6, pp. 331–346, 2003.

[3] F. Elisei, M. Odisio, G. Bailly, and P. Badin, "Creating and controlling video-realistic talking heads," in *Proceedings of the Workshop on Audio-Visual Speech Processing (AVSP '01)*, pp. 90–97, Aalborg, Denmark, 2001.

[4] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: driving visual speech with audio," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '97)*, pp. 353–360, Los Angeles, Calif, USA, August 1997.

[5] T. Ezzat and T. Poggio, "Visual speech synthesis by morphing visemes (MikeTalk)," Tech. Rep. A.I Memo No: 1658, MIT AI Lab, 1999.

[6] U. K. Goyal, A. Kapoor, and P. Kalra, "Text-to-audio visual speech synthesizer," in *Proceedings of the 2nd International Conference on Virtual Worlds*, pp. 256–269, July 2000.

[7] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '96)*, vol. 1, pp. 373–376, Atlanta, Ga, USA, May 1996.

[8] E. Cosatto and H. P. Graf, "Photo-realistic talking-heads from image samples," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 152–163, 2000.

[9] C. Weiss, "A framework for data-driven video-realistic audio-visual speech synthesis," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04)*, Lisbon, Portugal, May 2004.

[10] K. Liu and J. Ostermann, "Realistic facial animation system for interactive services," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech '08)*, Brisbane, Australia, September 2008.

[11] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '02)*, vol. 21, pp. 388–398, San Antonio, Tex, USA, July 2002.

[12] B. J. Theobald, J. A. Bangham, I. A. Matthews, and G. C. Cawley, "Nearvideorealistic synthetic talking faces: implementation and evaluation," *Speech Communication*, vol. 44, no. 1–4, pp. 127–140, 2004.

[13] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," in *Proceedings of the European Conference on Computer Vision*, vol. 2, pp. 484–498, 1998.

[14] K. W. Grant and S. Greenberg, "Speech intelligibility derived from asynchrounous processing of auditory-visual information," in *Proceedings of the Workshop on Audio-Visual Speech Processing*, pp. 132–137, 2001.

[15] K. W. Grant, V. Van Wassenhove, and D. Poeppel, "Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony," *Speech Communication*, vol. 44, no. 1–4, pp. 43–53, 2004.

[16] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[17] S. Fagel, "Joint audio-visual units selection—the javus speech synthesizer," in *Proceedings of the International Conference on Speech and Computer*, 2006.

[18] B.-J. Theobald, S. Fagel, G. Bailly, and F. Elisei, "LIPS2008: visual speech synthesis challenge," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech '08)*, Brisbane, Australia, September 2008.

[19] W. Mattheyses, W. Verhelst, and P. Verhoeve, "Robust pitch marking for prosodic modification of speech using td-psola," in *Proceedings of the 2nd Annual IEEE Benelux/DSP Valley Signal Processing Symposium (SPS-DARTS '06)*, pp. 43–46, Antwerp, Belgium, March 2006.

[20] L. Latacz, Y. Kong, and W. Verhelst, "Unit selection synthesis using long non-uniform units and phoneme identity matching," in *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, pp. 270–275, Bonn, Germany, August 2007.

[21] L. Latacz, Y. Kong, W. Mattheyses, and W. Verhelst, "An overview of the VUB entry for the 2008 blizzard challenge," in *Proceedings of the Interspeech Blizzard Challenge*, 2008.

[22] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, 1990.

[23] W. Mattheyses, L. Latacz, Y. Kong, and W. Verhelst, "A flemish voice for the nextens text-to-speech system," in *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference*, pp. 148–153, Lublijana, Slovenia, October 2006.

[24] G. Wolberg, *Digital Image Warping*, IEEE Computer Society Press, Los Alamitos, Calif, USA, 1990.

[25] E. Krahmer, S. Ruttkay, M. Swerts, and W. Wesselink, "Pitch, eyebrows and the perception of focus," in *Proceedings of the Speech Prosody*, pp. 443–446, Aix-en-Provence, France, April 2002.

[26] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, “Visual prosody: facial movements accompanying speech,” in *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 396–401, Washington, DC, USA, May 2002.

[27] A. Conkie and I. Isard, “Optimal coupling of diphones,” in *Proceedings of the 2nd ESCA/IEEE Workshop on Speech Synthesis*, pp. 293–304, 1994.

[28] W. Verhelst and M. Roelands, “An overlap-add technique based on waveform similarity (WSOLA) for high-quality time-scale modification of speech,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’93)*, pp. 554–557, Minneapolis, Minn, USA, April 1993.

[29] J. Kominek and A. W. Black, “The CMU arctic speech databases,” in *Proceedings of the 5th ISCA Speech Synthesis Workshop*, pp. 223–224, Pittsburgh, Pa, USA, 2004.