Multimodal Coherency Issues in Designing and Optimizing Audiovisual Speech Synthesis Techniques

Wesley Mattheyses, Lukas Latacz and Werner Verhelst

Vrije Universiteit Brussel
Dept. ETRO-DSSP, Brussels, Belgium
{wmatthey,llatacz,wverhels}@etro.vub.ac.be

Abstract

This paper proposes a 2D audiovisual text-to-speech synthesis system that constructs the output signal by selecting and concatenating multimodal segments containing natural combinations of audio and video. We describe the experiments that were conducted in order to assess the impact of this joint audio/video synthesis technique on the perceived quality of the synthetic speech. The experiments indicate that a maximal level of audiovisual coherence present in the output speech improves the perceived quality when compared to the traditional approach of synthesizing the visual signal separately from the audio. In addition, we measured that there is a same maximum allowable desynchronization between the audio and the image sequence, irrespective whether the degree of desynchronization is constant or time varying. This tolerance is used in the synthesizer for further optimizing the segment cuttings points in the audio and in the video mode.

Index Terms: audiovisual speech synthesis, multimodal unit selection, audiovisual synchrony

1 Introduction

The task of an audiovisual text-to-speech (TTS) synthesizer is to create an artificial speech signal, based on some input text. Where classical auditory TTS systems concern only the creation of the audible speech mode, audiovisual synthesizers also produce a talking head accompanying this auditory speech. Their functionality can be applied in various domains where machine-human communication is needed. For example, in e-commerce and e-learning environments, a virtual agent can be useful since the addition of the synthetic visual speech to the artificial audible speech will make the user feel more confident and attentive [1] and will consequently enhance the quality of the communication towards the customer or student. Audiovisual text-to-speech (AVTTS) systems can be roughly categorized in two groups: 3D model based synthesizers and 2D data-based systems. Whereas the 3D model based systems try to vary the polygons of a 3D mesh of the head in accordance with the auditory speech, 2D data based synthesis focuses on the creation of a photorealistic visual speech signal. This photorealistic output signal can be applied in a huge variety of applications since it is similar to standard 2D video and TV-broadcast.

2 Multimodal unit selection for AVTTS

The most common data-driven speech synthesis strategy is the unit-selection paradigm [2]. Based on the input text, the system's linguistic front-end creates a series of phonemic, linguistic and prosodic parameters, which are then used to select the most appropriate set of segments from a speech database provided to the system. This selection is based on target cost functions that indicate how well a segment matches the target speech, and join cost functions which indicate how well two consecutive segments can be concatenated without the creation of disturbing artifacts. After selection, these selected segments are concatenated to construct the final output signal. For a long time now it is known that there exists an important correlation between the audio and the video mode in multimodal speech perception, since perceptual effects like the McGurk effect [3] can drastically degrade this perception if the presented speech suffers from multimodal coherency issues. Nevertheless, almost all 2D AVTTS systems described in the literature, e.g. [4]-[6], synthesize the auditory and the visual output speech separately from each other. For each mode they use a different database and a different unit selection computation, after which both synthetic speech modes are synchronized and multiplexed into one final output signal. This strategy has the advantage that, while synthesizing one of the two modes, the selection of a certain speech unit only depends on the properties of that particular mode and is independent from the other mode. The separate selection of the two modes results in a huge amount of possible audio/video combinations to apply in the final output speech. Also, this strategy has the advantage that the size of the visual database, used for synthesizing the video track, can be kept limited in comparison to the auditory database, which is a huge benefit in terms of data storage. However, synthesizing both modes separately makes it hard to achieve a constant high level of intermodal coherence in the output speech, since the resulting signal will consist of artificial combinations of audio and video. Therefore, we designed a multimodal unit selection technique where the system selects and joins segments from an audiovisual database containing an original combination of auditory and visual speech. This approach, which was also mentioned in a preliminary study by Fagel [7], minimizes any mismatch between the output audio and the output video and it assures a maximal multimodal coherence in the final speech signal. The downside of this audiovisual selection strategy is obvious: selecting and concatenating an appropriate set of multimodal speech fragments is much harder than for the unimodal case. The target cost functions have to be defined in such a way that the selected fragments match the target speech as much as possible in both the auditory and the visual domain. Furthermore, audiovisual join costs have to be introduced to ensure a smooth concatenation in both modes. For more details on the multimodal selection technique applied

in our AVTTS system, the reader is referred to [8]. It is necessary to investigate the benefits of this coupled audio/video synthesis and to assess the importance of the synchrony and the coherency between the synthetic articulations in the audio and in the video mode of synthesized multimodal speech. In this paper we describe the experiments we conducted in order to evaluate the trade-off between freedom in selection/optimization and the minimization of audiovisual mismatches that can cause a degraded perception of the output signal.

3 The importance of audiovisual coherence for the perceived synthesis quality

3.1 Introduction

In a first series of experiments, we assessed the impact of the joint audio/video synthesis technique on the perception of the resulting synthetic visual speech. These experiments are described in detail in [9] and partly summarized below. Four different synthesis strategies were used for comparison. As audiovisual data we used the LIPS08 database [10], containing about 20 minutes of non-expressive English speech. A first group of samples (MUL) were synthesized using the multimodal selection and concatenation strategy discussed in the previous section, which means that these samples consist of concatenated original combinations of audio and video. The selection parameters were chosen such as to maximize the synthesis quality of the visual mode by increasing the weights of the visual join costs at the expense of the auditory join costs. The applied concatenation strategy desynchronized the original combinations of audio and video by maximum one video frame, which should be unnoticeable for a viewer. A second group of samples (SAV) were created by synthesizing both modes separately, using consecutively only the audio and only the video of the same audiovisual database. After synthesis, the synchrony between both modes was assured by performing a non-uniform time-scaling on the audio track using WSOLA [11] in order to align the audio with the video track. Thus, although 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. The third set of sample sentences (SVO) were created in the same way as the (SAV) set, but in this case we used the CMU ARCTIC database [12] of an English female speaker to create the synthetic audio mode. A fourth group (RES) contained samples from which the audio mode was directly extracted from a sentence contained in the LIPS08 database. This particular sentence was then excluded from the database and the video mode of the (RES) sample was synthesized using the same 'best video quality' settings as were used for the (MUL) samples and for the video track of the (SAV) and the (SVO) samples. Both modes were aligned and joined in the same way as we did for the (SVO) and (SAV) samples. Finally, an additional set of samples (ORI) was added, which contained original audiovisual sentences extracted from the audiovisual database used for synthesis.

3.2 Perception quality experiment

The main goal in this test was to investigate how the perception of the visual speech is influenced by the properties of the accompanying auditory speech. The participants were asked to rate the naturalness of the mouth movements displayed in the 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'. For this test all five sample sets described in section 3.1 were used. The results are summarized in table 1.

	ORI	MUL	SAV	SVO	RES
Mean	4.91	3.37	3.11	2.53	3.01
Median	5	3	3	2	3
ORI	-	2.88e-8	2.21e-8	1.77e-8	7.19e-9
MUL	-	-	0.210	0.00140	0.0474
SAV	-	-	-	0.0118	0.436
SVO	-	-	-	-	0.0296
RES	-	-	-	-	-

Table 1: *Mean, median and the significance levels resulting from a Wilcoxon signed-rank test*

3.3 Discussion

Note that for all but the (ORI) samples, the visual mode is synthesized by using the same database and the same 'best video quality' settings. This implies that any significant difference in perceived quality of the visual speech will be caused by the auditory speech played along with the visual mode. The most surprising result is the 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, we expected an optimal perception of these video tracks. However, the results indicate that the viewers gave a higher rating to the samples of the (MUL) group. This shows that for a quality perception of the naturalness of the visual speech signal, a high amount of audiovisual coherence is more important than the individual quality of the auditory speech. Furthermore, by comparing the (MUL) results to the (SVO) results, a clear preference for the (MUL) samples in noticeable. This can again be explained by the fact that the audiovisual coherence of the (MUL) samples is much better than found in the (SVO) samples. It is also worth mentioning that by comparing the (SVO) to the (RES) samples, there is an indication that a 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 impact is only secondary.

4 Optimal joining of multimodal speech fragments

4.1 Introduction

High quality concatenation of the selected multimodal segments should result in an audiovisual speech track which exhibits smooth and natural variations in both its auditory and its visual mode. Furthermore, the concatenation of the original combinations of audio and video should not affect their multimodal coherence. In other words, at each time instant the signal in the audio track of the joined multimodal speech should be accompanied by its naturally matching visual signal. In this section we describe some possible approaches to tackle the multimodal concatenation problem and we discuss the experiments we conducted in order to assess their impact on the perceived output quality.

4.2 Join algorithms

In order to join two audiovisual segments, two concatenation actions are needed: one for joining the waveforms of the audio tracks and one for joining the video chunks of the visual tracks. In our 2D AVTTS synthesis, the concatenation in the audio mode is tackled by a pitch-synchronous crossfade technique, as described in [13]. This technique minimizes the introduction of anomalous pitch periods when two voiced speech signals of different pitch levels are joined. The joining of the two video chunks is achieved by creating intermediate frames using the mesh-warping technique. A detailed description and examples of this join strategy are given in [8]. Summarized, the system first detects for each frame a set of landmark points which indicate the location of the lips, the eyes, the nose and the face. These landmark points are then used as feature primitives for the mesh-warping algorithm, which creates a series of intermediate frames to morph the mouth instance found at the end of the first segment into the mouth instance present at the beginning of the second segment.

4.3 Optimal coupling

When two audiovisual segments are to be joined, an audio sample and a video frame have to be selected in the first and in the second segment to use as endpoint and as startpoint, respectively. These points are referred to as cutpoints: the time instances on which the two segments are cut from the database. For each join, the system first selects an optimal pair of cutpoints in the audio track, based on the minimization of the auditory join cost for the particular concatenation, as described in [13]. Afterwards, the system needs to select the cutpoints in the visual mode so the video chunks can be concatenated. Two different approaches were implemented and evaluated. In a first approach the system tries to minimize the audiovisual asynchrony by selecting the join-frames as close as possible to the audio cutpoints. A different technique consists in optimizing each video concatenation by tolerating some variations between the cutpoints in the video and the cutpoints in the audio. However, this will cause extra multimodal incoherencies and asynchronies of which the impact needs to be investigated.

4.3.1 Approach 1

A first important observation is that the sample rate of the audio signal (44100Hz) is much higher than the sample rate of the video signal (25Hz). This implies that locating the ideal join position in the audio track can be performed on a much smaller scale than for cutpoint in the video track. The most straightforward technique to determine the video join position is to select the video frame that lies on the time-axis the closest to the audio cutpoint. This will cause a minimal desynchronization between the audio track and the video track when they are added to the sequence of already joined segments. However, even the smallest difference between the cutpoints in both modes causes a discrepancy between the length of the audio track and the length of the video track of the selected multimodal segment. This will result in a desynchronization between the two modes of the audiovisual segment that will be joined to the output speech after this particular one. To calculate the resulting asynchrony ($DES(i)$) for a certain segment i, we have to take into account the difference in audio/video startpoint for the particular segment ($\Delta SP(i)$) and the difference in length of the already joined audio track and the already joined video track to which the current segment will be added ($\Delta L(i-1)$):

$$DES(i) = \Delta SP(i) + \Delta L(i-1) \quad with \tag{1}$$

$$\Delta SP(i) = t_{videocut}(i) - t_{audiocut}(i) \tag{2}$$

$$\Delta L(i-1) = \sum_{n=1}^{i-1} audiolength(n) - \sum_{n=1}^{i-1} videolength(n) \tag{3}$$

The effect of a uniform asynchrony between both modes in audiovisual speech perception is well documented in the literature [14]. The most important conclusion is that humans are highly sensitive to a lead of the audio track in front of the video track, but there exists quite a tolerance on the lead of the visual speech in front of the auditory speech. We can exploit this property to cope with the unavoidable asynchrony caused by the difference between the cutpoints in both modes: after calculating the best audio cutpoint, the join position in the video track is determined by selecting the video frame that is closest to the audio join position and that introduces an audiovisual asynchrony between zero and one video frame lead of the visual speech, which is calculated using equation (1).

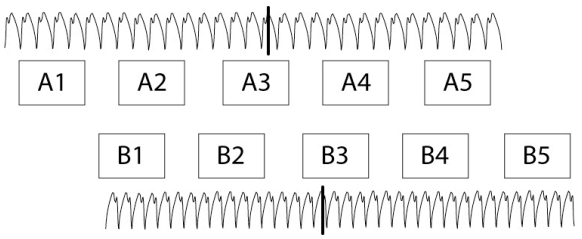

Figure 1: *Optimal video coupling*. This figure displays two audiovisual signals that are to be joined. The cutpoints in the audio tracks are indicated. The most straightforward strategy is to select A3-B3 as join position since this minimizes the audiovisual asynchrony: it is kept between zero and one video frame lead. A possible optimal coupling technique is to determine for both signals a set of candidate frames A1-A5 and B1-B5, from which the most optimal pair is selected. Note that all other combinations than A1-B1, A2-B2, etc. will result in an increased audiovisual desynchronization.

4.3.2 Approach 2

Similarly to the optimal coupling technique applied for the audio joins, the smoothness of the video mode can be enhanced by fine-tuning the video cutpoints at each concatenation. Therefore, a set of possible endframes and a set of possible startframes are selected in the neighborhood of the join positions in the audio mode. Afterwards, one frame from each set is chosen as final cutpoint, based on the minimization of the visual join cost calculated for every combination of endframe-startframe (see figure 1). Note that this technique will cause extra desynchronizations between the joined audio and the joined video track, since there will be an increased and varying difference between the video cutpoints and the audio cutpoints used at each concatenation. This results in an audiovisual output signal containing an asynchrony between its two modes that varies from segment to segment. This optimal coupling algorithm can be tweaked by three parameters: the maximal local audio lead (referred to as negative desync), the maxi-

mal local video lead (referred to as positive desync) and a search-length parameter which defines the size of each set of candidate join frames. This search-length defines the amount of asynchrony that can be introduced by a single concatenation, while the other two parameters define the allowed limits of the combined desynchronization effect of consecutive concatenations. This implies that the set of possible join frames will in general not be centered around the audio cutpoint, since the absolute value of the minimal and maximal allowed asynchrony will be different and also because the contribution of the previous concatenations should be taken into account (equation (1)). In the next section we will describe the experiments we conducted in order to assess the effects of this optimal video coupling technique on the perceived audiovisual synchrony and on the measured and perceived smoothness of the synthetic visual speech.

4.4 Perception of local audiovisual desynchronization

4.4.1 Subjective test

Literature about the effects of a uniform and constant audiovisual desynchronization on the perception of audiovisual speech mentions $-50ms$ and $+200ms$ as tolerable bounds for the asynchrony without being noticed by the viewers [14]. However, in our AVTTS synthesis the introduced asynchrony between the audio and the video mode will be non-uniform, as it varies among every added segment. To determine an optimal setting for the minimal and maximal desynchronization parameters described in section 4.3.2, we conducted a listening experiment were different settings of these parameters are evaluated. Note that the actual asynchrony resulting from the optimal coupling technique will not always reach the maximal allowed levels, since this depends on which candidate join frames are selected to minimize the join cost. Thus, the most convenient strategy is to search for the most extreme values of local desynchronizations that are unnoticed by the participants and to use these values as future parameter settings. For the listening test we synthesized 5 different sentences (with mean word count of 17 words) using the LIPS08 database and various parameter values. Based on the actual resulting desynchronizations, we selected 25 samples to cover the range of asynchronies we want to evaluate. The participants were shown pairs of sample syntheses of a same sentence, where each time one of the two samples was an optimized one (thus containing some varying local asynchrony, see section 4.3.2) and the other a non-optimized one (thus containing no noticeable asynchronies, see section 4.3.1). The participants were asked to write down which of the two samples scored best in terms of audiovisual synchrony. If they could not notice any difference, they were asked to answer 'no difference'. 7 people participated in the test, 3 of which were experienced in speech processing. In table 2 the results of the experiment are summarized. To obtain the values shown in table 2 we grouped the samples in accordance with their resulting minimal and maximal local asynchrony. It shows a detection ratio of less than 20% for the samples where the audio lead is smaller than 0.04s and the video lead is smaller than 0.2s. These parameters settings will be applied for the second listening test described later on in this paper. Note that the test set of 25 samples is insufficient to deliver several samples for each group, which makes it impossible to define the exact thresholds for noticing non-uniform audiovisual asynchronies. Nevertheless, the results obtained in the test are sufficient to determine suitable parameter values for the optimal coupling technique. Furthermore, it can be seen as a preliminary study on the effects of a time-varying multimodal desynchronization on audiovisual speech perception. The results indicate that the thresholds for noticing non-uniform audiovisual asynchronies will be quite similar to the values found for uniform desynchronizations. It seems to be the case that the length of a multimodal asynchrony occurring in audiovisual speech communication has no influence on the perception: even very short mismatches (occuring when short segments are selected) between the audio and the video are noticed. This result is in agreement with earlier experiments on audiovisual perceptual effects (e.g.: the McGurk effect [3]), where it was found that even coherence mismatches lasting only the duration of a single phoneme could drastically decrease the intelligibility and/or the perceived quality of the multimodal speech signal.

	Min desync			
Max desync	0s	-0.04s	-0.08s	-0.15s
0s		0%	26%	90%
0.1s	0%	0%	10%	100%
0.2s	15%	0%	20%	60%
0.4s	46%	no samples	40%	100%

Table 2: *Percentage of noticed desynchronizations.*

4.4.2 Tweaking the optimal video coupling

The results obtained indicate that the introduction of a varying audiovisual asynchrony often results in a poorer perception quality. Perhaps it would be safer to adapt the optimal coupling strategy in such a way that no extra asynchrony is introduced. This can be realized by selecting only those pairs of candidate join frames that introduce no change in signal length: for the example given in figure 1 this implies that only pairs A1-B1, A2-B2, etc. are considered. The downside is that the amount of possible combinations endframe/startframe will be much more limited using this technique, probably resulting in less optimized concatenations. For this strategy there exists only a single parameter to tweak the optimization, namely the search-length that determines the amount of candidate endframes and startframes. The higher this value, the more possibilities there are to smooth the concatenation. However, choosing a high value could result in a degrading of the multimodal coherence, since at the join instants artificial combinations of audio and video will occur.

4.5 Objective smoothness measures

To assess the effect of the optimal coupling strategies on the smoothness of the concatenated visual speech signal, an objective test was performed. For each synthesized sentence we created new metadata by tracking the facial landmarks and the PCA coefficients of the mouth area through the movie. A smoothness measure is defined as the linear combination of the summed Euclidean differences of the newly tracked mouth landmark positions and the Euclidean difference of the newly created PCA coefficients for every two consecutive frames in the final output speech located at the transition points between the selected segments. By summing these measures and by dividing this value by the amount of sampled differences, we get a single measure for each synthesized sentence. For this test we synthesized 11 different sentences (mean word count = 15 words) using the LIPS08 database and 6

different settings, summarized in table 3. The results obtained are shown in figure 2 by means of a box plot.

Group	Method	Searchlength	Min Des	Max Des
I	None	-	-	-
II	A	0.20s	-0.04s	0.20s
III	A	0.20s	-0.05s	0.35s
IV	B	0.08s	-	-
V	B	0.20s	-	-
VI	B	0.40s	-	-

Table 3: *Optimal coupling settings* All sentences were synthesized using these 6 settings. For group I, no optimal coupling was used (see section 4.3.1). For groups II to VI, at each concatenation different combinations endframe/startframe were considered, where in method A every possible combination was allowed (see section 4.3.2) and in method B only those pairs that keep the length of the joined video track unaltered were chosen (see section 4.4.2).

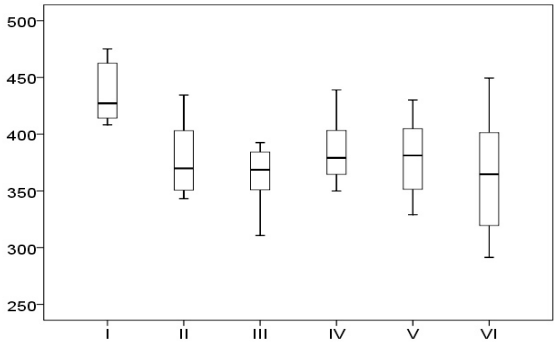

Figure 2: *Measured mean transition costs*

Figure 2 shows that the non-optimized samples result in worse measures than all optimized groups (this was proved by paired sample t-tests) and that the visual concatenations are indeed smoothed by the optimal coupling technique. Another important result is that there is only little difference between method A and method B, despite the fact that for a same search-length value method A takes much more possible combinations of endframe/startframe into account. A last observation is the noticeable improvement of the smoothness measure if the search-length and/or the tolerable limits of asynchrony are increased.

4.6 Subjective test

4.6.1 Motivation and description

To assess the effects of the optimal video coupling techniques on the perception of the synthetic visual speech, a subjective listening test was performed. From the set of samples used in the objective test described above, we took for each of the 11 sentences the samples from the groups I, II and V. From each sentence, two pairs of samples were played to the participants. Each pair contained the synthesis from group I and a synthesis from group II or from group V. The viewers were asked to write down which visual speech track they preferred. They were told to pay attention especially to the smoothness of the mouth movements and to the audiovisual coherence. If they had no preference, they were asked to answer 'no difference'. 9 people participated in this test, 6 of them being experienced in speech processing.

4.6.2 Results and discussion

The summarized preference results are given in table 4. After converting the obtained data to $[-1, 0, 1]$ rating scores, we performed a Wilcoxon Signed Rank Test for both comparisons (see also table 4).

Test	Preference	Amount
Group I - Group II	Group I > Group II	38
	Group I = Group II	39
Total = 99	Group I < Group II	22
Group I - Group V	Group I > Group V	19
	Group I = Group V	60
Total = 99	Group I < Group V	20

Test	Asymp. significance
Group I - Group II	0.039
Group I - Group V	0.87

Table 4: *Subjective test results*

The results indicate that the participants showed no preference for the optimized samples. For the comparison between groups I and V, the majority of the answers is 'no difference' and both groups are rated equally. On the other hand, there is a noticeable tendency that viewers disliked the samples from group II in favor of the non-optimized syntheses. Inspection of the reported answers and feedback from the participants point out several explanations for these results. Firstly, the answers differ heavily among the users: some of them tend to generally like or dislike the optimized samples and some participants answered very often 'no difference' while others reported much more preferences. Furthermore, almost all users informed us that they often did notice an improved smoothness of a certain visual signal, but that many times those sentences contained audiovisual coherence issues which often forced them to report a preference for the non-optimized synthesis. Probably this observation explains why especially the samples from group II were disliked: these contained a varying multimodal asynchrony caused by the optimal coupling technique. In contrast to the results obtained in the first subjective test (see table 2), the parameter settings used are apparently still too high to minimize any noticeable desynchronization. On the other hand, it seems that the smoothing effect applied to the samples of group V is not easily noticed by the users. Even more, some of the collected answers indicate that this smoothing strategy sometimes causes a decrease in perceived quality. This can be explained by the fact that this alternative optimal coupling technique does not introduce any asynchrony, however at the join positions unnatural combinations of audio and video can occur (e.g.: the selection of frames A1-B1 or A5-B5 displayed in figure 1). Apparently, the effect of these small audiovisual mismatches can be large enough for a user to dislike the smoothed signal over the non-optimized one.

5 Conclusion

In this paper we studied the impact of the level of audiovisual coherence on the perceived quality of AVTTS synthesis. In a

first experiment, the impact of a joint audio/video selection was evaluated. The results show that the quality of the synthetic visual speech is rated highest if the multimodal output signal contains a maximal audiovisual coherence: the samples synthesized using the multimodal selection strategy were given better ratings than the samples created by combining different unimodal syntheses. Furthermore, they were even rated slightly better than the samples where the audio mode consisted of natural speech. In addition, we studied the audiovisual concatenation problem, where we investigated some techniques to smoothen the visual signal by optimizing the video joins. This optimization causes time-varying local asynchronies and/or incoherencies between both output modes. Results from a subjective test indicate that earlier published values for just noticeable audiovisual asynchrony hold in the non-uniform case as well (i.e., they hold for both constant and time varying asynchrony). A possible explanation for this resides in the fact that human speech perception is for a great deal based on predictions: by observing natural speech communication we learned 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. Since varying audiovisual desynchronizations do not exists in natural speech signals, it can be expected that there is no such thing as a 'temporal window' in which we are less sensitive to the audiovisual asynchrony in multimodal speech perception. Furthermore, objective measurements proved that the optimization of the visual joins indeed improves the smoothness of the synthetic visual speech. However, a subjective listening test conducted to assess the effects of the optimal coupling on the perception of the audiovisual speech showed no indication that users preferred the smoothed synthesis samples over the non-optimized ones. Quite often the participants did notice an increased smoothness of the visual track, but they detected that this results in a decrease in multimodal coherence between the audio and the video mode. The optimal coupling caused some sort of disturbing under-articulation effect: rapid variations in the audio mode are not seen the video mode. These findings are in line with the results from the first experiment: in order to attain a high-quality perception of the audiovisual signal, a maximal multimodal coherence is crucial. The avoidance of any mismatch between both output modes seems to be as important as the individual optimization a certain mode. This implies that an optimized multimodal unit selection, where the amount of candidate units is increased by decoupling the original combinations audio/video, will only be advantageous if we find methods to increase the compatibility between the separate audio and video. Similarly, a technique to improve the audiovisual concatenations should be designed in a truly audiovisual way, where the interpolation is optimized for the combined audiovisual information in stead of applying two separate optimizations like in the techniques discussed in this paper. Sample syntheses of the AVTTS system can be found at: www.etro.vub.ac.be/Research/DSSP/Projects/avtts/demo_avtts.htm.

6 Acknowledgements

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.

References

[1] Pandzic, I., Ostermann J. and Millen D., "Users Evaluation: Synthetic talking faces for interactive services", The Visual Computer, Volume 15 2330-2340, 1999

[2] Hunt, A. and Black, A., "Unit selection in a concatenative speech synthesis system using a large speech database", International Conference on Acoustics, Speech and Signal Processing, 373–376, 1996

[3] McGurk, H. and MacDonald, J., "Hearing lips and seeing voices", Nature, Volume 264 746–748, 1976

[4] Ezzat, T., Geiger, G. and Poggio, T., "Trainable videorealistic speech animation", Association for Computing Machinery's Special Interest Group on Graphics and Interactive Techniques, Volume 21 388–398, 2002

[5] 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, 619–622, 2000

[6] Theobald, B.J., Bangham, J.A., Matthews, I.A. and Cawley, G.C., "Near-videorealistic synthetic talking faces: implementation and evaluation", Speech Communication, Volume 44 127–140, 2004

[7] Fagel, S., "Joint Audio-Visual Units Selection - The Javus Speech Synthesizer", International Conference on Speech and Computer, 2006

[8] Mattheyses, W., Latacz, L., Verhelst, W. and Sahli, H., "Multimodal Unit Selection for 2D Audiovisual Text-to-Speech Synthesis", Springer Lecture Notes in Computer Science, Volume 4261 125–136, 2008

[9] Mattheyses, W., Latacz, L. and Verhelst, W., "On the importance of audiovisual coherence for the perceived quality of synthesized visual speech", EURASIP Journal on Audio, Speech, and Music Processing, SI: Animating Virtual Speakers or Singers from Audio: Lip-Synching Facial Animation, 2009

[10] Theobald, B.-J., Fagel, S., Bailly, G. and Elisei, F., "LIPS2008: Visual speech synthesis challenge", Interspeech 2008, 1875–1878, 2008

[11] Verhelst, W. and Roelands, M., "An Overlap-Add Technique based on Waveform Similarity (WSOLA) for High-Quality Time-Scale Modification of Speech" International Conference on Acoustics, Speech, and Signal Processing, 554–557, 1993

[12] Kominek, J. and Black, A.W., "The CMU Arctic speech databases", 5th ISCA Speech Synthesis Workshop, 223–224 2004

[13] 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

[14] Grant, K.W., Van Wassenhove, V. and Poeppel, D., "Detection of auditory (cross-spectral) and auditoryvisual (cross-modal) synchrony", Speech Communication, Volume 44:Special Issue on Audiovisual Speech Processing 43–53, 2004