Towards a True Acoustic-Visual Speech Synthesis

Asterios Toutios, Utpala Musti, Slim Ouni, Vincent Colotte, Brigitte Wrobel-Dautcourt, Marie-Odile Berger

Towards a True Acoustic-Visual Speech Synthesis

Asterios Toutios, Utpala Musti, Slim Ouni, Vincent Colotte,
Brigitte Wrobel-Dautcourt, Marie-Odile Berger

LORIA, UMR 7503, BP 239, 54506 Vandœuvre-lès-Nancy, France
{toutiosa,musti,slim,colotte,wrobel,berger}@loria.fr

Abstract

This paper presents an initial bimodal acoustic-visual synthesis system able to generate concurrently the speech signal and a 3D animation of the speaker's face. This is done by concatenating bimodal diphone units that consist of both acoustic and visual information. The latter is acquired using a stereovision technique. The proposed method addresses the problems of asynchrony and incoherence inherent in classic approaches to audiovisual synthesis. Unit selection is based on classic target and join costs from acoustic-only synthesis, which are augmented with a visual join cost. Preliminary results indicate the benefits of this approach, since both the synthesized speech signal and the face animation are of good quality.

Index Terms: audiovisual speech synthesis, talking head, bimodal unit concatenation, diphones

1. Introduction

During the last decade, interest in audiovisual speech increased and more studies exist in both perception and synthesis. Speech is no longer considered as purely acoustic but it is considered as a bimodal means of communication. The first modality is audio, provided by the acoustic speech signal, and the second is visual, provided by the face of the speaker. Actually, the speech signal is the acoustic consequence of the deformation of the vocal tract under the effect of the movements of articulators such as the jaw, lips, and tongue. Moreover, there is more and more research showing the existence of a clear correlation between the face and the vocal tract. Thus, it is quite natural to find out that acoustics and face movements are correlated [1, 2].

Research in audiovisual speech intelligibility showed the importance of the information provided by the face especially when the audio is degraded [3, 4, 5]. Moreover, in [4] the authors showed that when the acoustic is degraded or missing, the natural face provides two thirds of the missing auditory intelligibility, the face without the inner mouth (without the tongue) provides half of the missing intelligibility and the lips restores third of it. This indicates that audiovisual synthesis should pay careful attention to model the part of the face that participates actively during speech, i.e., mainly the lips and lower part of the face.

In the vast majority of recent works, data-driven audiovisual speech synthesis, i.e., the generation of face animation together with the corresponding acoustic speech, is still considered as the synchronization of two independent sources: synthesized acoustic speech (or natural speech aligned with text) and the face animation [6]. However, achieving perfect synchronization between these two streams is not straightforward and presents several challenges related to audio-visual intelligibility. Furthermore, it can be the case that the auditory and the visual information originate from different parts of the corpus, which may cause perceptual incoherence [7].

To avoid such problems, we propose, within the ViSAC project, to achieve synthesis with its acoustic and visible components simultaneously. Therefore, we consider a bimodal signal as one signal with two channels: acoustic and visual. This bimodality is kept during the whole synthesis process. The setup is similar to a typical concatenative (acoustic-only) speech synthesis setup, with the difference that here, the units to be concatenated consist of visual information alongside acoustic information. The concatenation unit adopted in our work is the diphone. The advantage of choosing diphones is that the major part of coarticulation phenomena is captured locally in the middle of the unit and the concatenation is made at the boundaries, which are acoustically and visually steadier. Actually, this choice is in accordance with current practices in concatenative speech synthesis.

The idea of concatenating bimodal units is not entirely new. First studies appeared in [8] and [9]. More recently, two advanced systems based on 2D images were presented in [10] and [7]. These works share several common characteristics with ours. Nevertheless, the combination of features of our system, as presented in this paper, is unique. In fact, our particular ability to acquire and process large amount of parallel audiovisual data will be very important to keep improving the quality of our results. In addition, a good coverage of the lower face by 3D markers is an important characteristic of our acquisition technique (see Fig. 1). In fact, the visual information is 3D data, pertaining to the movements of a large number of markers painted on the face of the speaker. The number of markers is large enough to allow accurate reconstruction of the lips and all the area around them, which is important mainly because of their importance in audiovisual intelligibility. As a matter of fact, one of our goals is to be able to accurately animate the lips and thus the articulation and the coarticulation reproduce similar behavior as that of the real speaker.

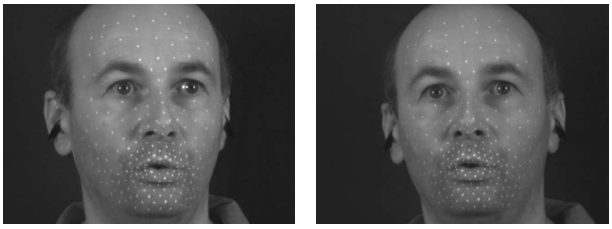

Figure 1: *A pair of stereovision images, showing the markers painted on the speaker's face.*

We expect a two-fold benefit from our approach. On the one hand, taking into account visual information in text-to-acoustic-

speech synthesis can improve the quality of speech synthesis by offering a more relevant distance measure for unit selection and concatenation. On the other hand, audiovisual synthesis can be improved due to the intrinsic consistency of combined acoustic-visual information.

In this paper, we first present our acquisition system and the data modeling. We used Principal Component Analysis (PCA) to reduce the dimensionality of the data. Second, the main idea of the bimodal selection and concatenation is presented. Then some preliminary results are presented to illustrate the potential benefits of our proposed method of combining acoustic and visual constraints in a concatenative system. Finally, we provide some comments and concluding remarks.

2. Data acquisition and modeling

2.1. Acquisition

Visual data acquisition was performed simultaneously with acoustic data recording, using an improved version of a low-cost 3D facial data acquisition infrastructure we developed in the past [11]. The system uses two fast monochrome cameras, a PC, and painted markers, and provides a sufficiently fast acquisition rate to enable an efficient temporal tracking of 3D points. The large majority of markers are detected by low-level processing of the stereo image pairs. However, there are also cases when markers cannot be directly detected, for example markers on the temples which may disappear when the speaker moves his/her head, or markers on the lips that are occluded during protrusion or closing of the mouth. In such cases the positions of the markers are estimated using an interpolation scheme that involves an initial 3D mesh of the face. Processing the data is still a lengthy work though (several weeks for 25 minutes of data).

The recorded corpus consisted of the 3D positions of 252 markers covering the whole face. However, the lower face was covered by 70% of all the markers (178 markers), where 52 markers were covering only the lips. We made this choice, to be able to capture the lip movement accurately and to be able to model finely the lips. The sampling rate was 188.27 Hz. The corpus was made of 319 medium-sized French sentences, covering about 25 minutes of speech, uttered by a native male speaker. A few extra sentences were recorded for testing purposes. These data were sub-sampled to 100 Hz, for easier labeling and alignment with acoustic parameters. In combination with the sub-sampling, the data were filtered using a low-pass filter with a cutoff frequency of 25 Hz. Such a processing removes additive noise from the visual trajectories without suppressing important position information. The speech signal was recorded at 16 kHz with 16-bit precision.

2.2. Principal Components

We applied principal component analysis on a subset of markers at the lower part of the face (jaw, lips, and cheeks–see Fig. 2). The reason for this choice was that the movements of markers on the lower part of the face are tightly connected to speech gestures, while markers on the upper part of the face either do not move, or their movements are of no direct relevance to speech. We retained the 12 first principal components, which explain about 94% of the variance of the lower part of the face.

These 12 components are shown in Fig. 3. The first two components, which explain 79.6% of the lower face variance, both account for combined jaw opening and lip protrusion gestures. For the first component, as the jaw closes, lips protrude.

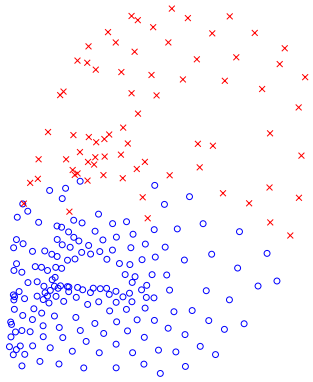

Figure 2: *PCA is applied on the information of 178 out of 252 painted markers (plotted in blue circles). The remaining of the markers (plotted in red crosses) are modeled only by their mean value, since they do not reflect explicit speech gestures.*

The effect is reversed for the second component: as the jaw *opens*, lips protrude. That is, the two basic gestures of jaw opening and protrusion appear not in isolation but intertwined. The third component accounts for lip opening, after removal of the jaw contribution. It is in good agreement with the lip opening factor typically described in articulatory models, as in Maeda's model [12], for instance. Some of the components are related to speech but augmented by some gestures are specific to speaker facial expression. For instance, components 4 and 5 capture lip spreading, however due to some asymmetry of our speaker's articulation, lip spreading is divided in two modes: one accounting for spreading toward the left side of the lips and one for spreading toward the right side. Component 6 is a smiling gesture, however it is not clear whether it is related to speech or pure facial expression. Components 7 to 12 seem to account for very subtle lip deformations, which we believe are idiosyncratic characteristics of our speaker.

Several experiments indicated that retaining as less as three components could lead to an animation which would be acceptable, in the sense that it would capture the basic speech gestures and would filter out almost all the speaker specific gestures. However, such an animation would lack some naturalness, which is mostly captured by secondary components. We are also in favor of keeping the specificity of the speaker gestures. Retaining 12 components leads to animations that are natural enough for all purposes.

One of the goals of our proposed system is to synthesize trajectories corresponding to the PCA-reduced visual information, for these 12 components, alongside the synthesized speech signal. The visual information of the lower face can be reconstructed using these 12 trajectories. The mean values of the positions of the markers at the upper part of the face may then be added to complete the face visualization.

3. Bimodal selection and concatenation

As in typical concatenative speech synthesis, a corpus was phonetized, analyzed linguistically, and partitioned into diphones. This corpus included information on position, duration, acoustic, visual (that is, the PCA-reduced representation) and linguistic parameters for each diphone. This corpus consisted of the 319 sentences. The size of this corpus is large enough compared to other works on audiovisual synthesis, but small

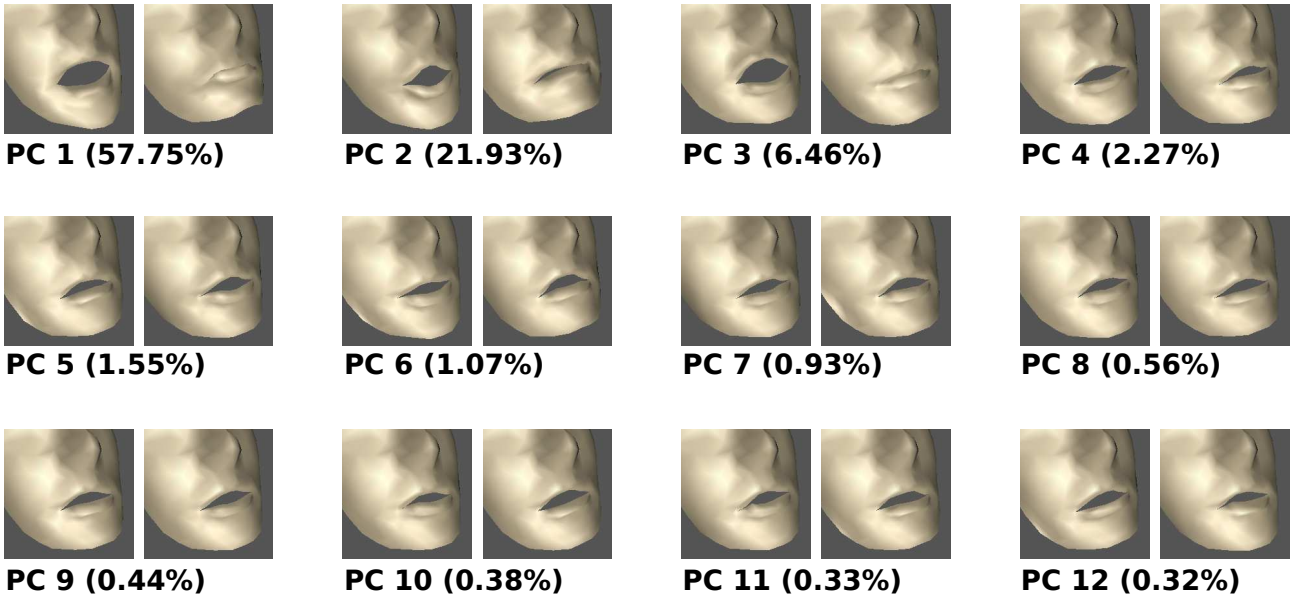

Figure 3: *The 12 first principal components of the facial data and the percentage of variance each of them explains. Each pair of images shows the deformation of the face when the corresponding component assumes a value of -3 (left) or +3 (right) standard deviations.*

compared to works on text-to-speech synthesis. The main goal of using this corpus is to study the feasibility of our technique. During this project, we are designing a larger corpus of acoustic and visual data, composed of about 2000 sentences (more than 2 hours of speech). The new corpus will cover at least all diphones of the French language and will contain several representations of these diphones in different contexts. Special care will be taken to account for visual variability alongside acoustic variability.

The input to the synthesizer system is a text that is automatically phonetized and partitioned into diphones. For each diphone, all possible candidates from the corpus must have the same phonemic label. A special algorithm is available to handle cases when there are no instance of the same diphone. The selection among these candidates is performed using the Viterbi algorithm. The result of the selection is the path in the lattice of candidates which minimizes a weighted linear combination of three costs: the target cost (TC), the acoustic join cost (JC), and the visual join cost (VC), that is

$$C = w_{tc}TC + w_{jc}JC + w_{vc}VC \quad (1)$$

where w_{tc}, w_{jc} and w_{vc} are weights to be chosen empirically by the experimenter.

The target cost is calculated on the basis of the linguistic analysis of the target utterance and is a weighted summation of the difference between the features of the candidate diphone and the features of the target diphone. Some of the features used are: syllable number and position in word, rhythmic group, and sentence; word number and position in rhythmic group and sentence; proximity of pauses; phoneme voicing, place and manner of articulation. The acoustic join cost is defined as the acoustic distance between the units to be concatenated, and is calculated using acoustic features at the boundaries of the units to be concatenated: fundamental frequency, spectrum, energy, and duration. For more details on the calculation of these costs see [13].

Similarly, the visual join cost is defined as the visual distance between the units to be concatenated. This is calculated

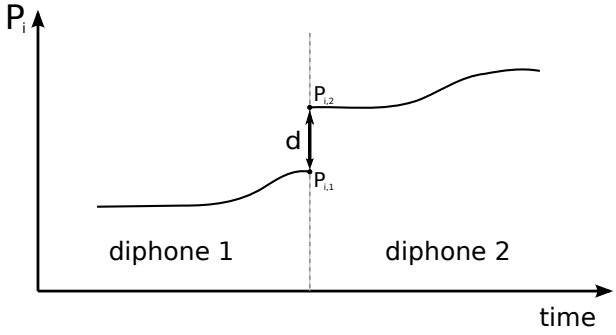

Figure 4: *Illustration of the visual cost calculation. The purpose is to minimize the distance d between the points $P_{i,1}$ and $P_{i,2}$ at the boundary of the two concatenated diphones.*

using the PCA transformed visual information at the boundaries of the units to be concatenated. That is:

$$VC = \sum_{i=1}^{12} w_i \left(P_{i,1} - P_{i,2}\right)^2 \quad (2)$$

where $P_{i,1}$ and $P_{i,2}$ are the values of the projection on principal component i at the boundary between the two diphones (see Fig. 4). The choice of weights w_i is generally up to the experimenter, however it should reflect the relative importance of the components. In accordance with [14], we chose these weights to be proportional to the eigenvalues of PCA analysis.

The selected diphone sequence is concatenated acoustically using a traditional technique, where pitch values are used to improve the join of diphones. At the moment, we do not apply any processing at all to the concatenated visual trajectories. Such a processing would be helpful in imposing continuity upon visual trajectories, at boundaries between diphones, especially for less important principal components. For the most important components, the trajectories are already acceptably continuous

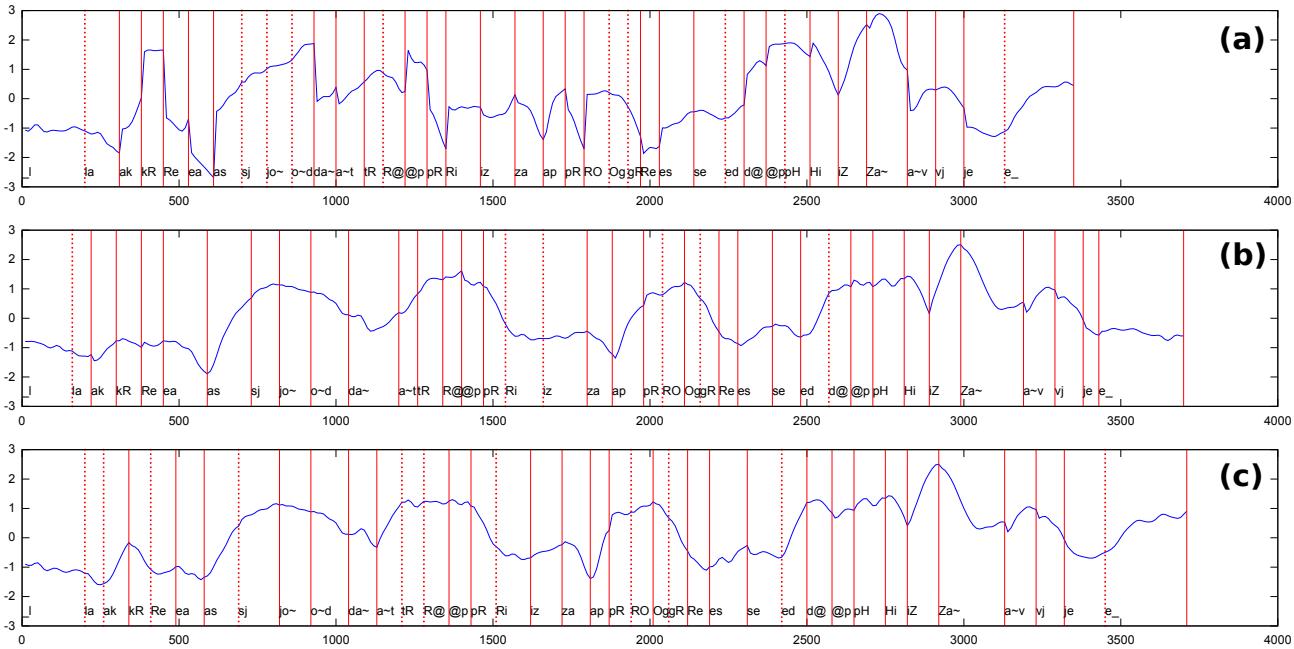

Figure 5: *Concatenated first visual principal component (in z-scored units) for the test sentence 'La création de l'entreprise a progressé depuis Janvier" when: (a) only target and acoustic join costs are minimized; (b) only visual join cost is minimized; (c) a weighted sum of all three costs is minimized. Horizontal axes denote time in milliseconds. The boundaries between diphones are marked. Dashed lines indicate that the combination of the two diphones exists consecutively in the corpus and is extracted "as is" from it, solid lines otherwise. SAMPA labels for diphones are shown.*

(given our choice of relative weights for the calculation of the visual cost), and we expect that the quality will be improved further with a larger corpus as more candidates will be present.

4. Results

Some examples of acoustic-visual synthesis are available at (`http://visac.loria.fr/demos`). To provide a preliminary evaluation of our system, we studied three cases for the weights of Eq. (1). First, we selected $w_{vc} = 0$, i.e., we did not apply a visual cost, and the system was driven only by linguistic and acoustic constraints, via target and acoustic join cost, respectively (acoustic-only case). The weights w_{tc} and w_{jc} were set to optimal values according to our experience with acoustic-only synthesis.

For the second case, target and acoustic join cost weights were set to zero. The system was driven only by visual information (visual-only case).

The third case was truly bimodal. The target and acoustic join cost weights were selected as in the first case. The visual weight was given such a value so that the contribution of the third (visual) term of Eq. (1) was roughly equal to the contribution of the first two terms.

In Fig. 5 we show the trajectory of the first principal component for the synthesis of the sentence: "La création de l'entreprise a progressé depuis Janvier". In the acoustic-only case (a) there are some obvious discontinuities in the visual trajectory, that results in visible jerks during the animation of the face. On the contrary, in the visual-only (b) and bimodal (c) cases the resulting visual trajectories are smoother. This is true for the first three components, however, as we move to less important components discontinuities start to appear gradually, as can be seen in Fig. 6. This is because the weights we chose in Eq. (2) (i.e. the eigenvalues) put a lot of emphasis on the first few components. Nevertheless, more important a principal component, more sensitive the animation is to discontinuities in the corresponding projection.

Regarding speech acoustics, the bimodal case results to a waveform that is much closer to the waveform resulting from the acoustic-only case, than to the one resulting from the visual-only case. Listening showed that the visual-only result, while intelligible, has several problems regarding phoneme duration, intonation and some audible discontinuities at boundaries between diphones.

Broadly, using in conjunction acoustic/linguistic and visual constraints combines benefits from using in isolation either acoustic/linguistic or visual constraints. Fig. 7 shows an animated sequence using all constraints for a part of the sentence shown in Figs. 5. The faces (shown only in part, to emphasize the mouth area) are represented here by sparse meshes. These will be mapped to meshes of much higher resolution later on which should add more realism to our final result [15].

5. Conclusion

We presented a bimodal acoustic-visual synthesis where synthesis is achieved simultaneously for both acoustic and visual channels by concatenating bimodal diphones. The system combines costs from acoustic-only speech synthesis with a visual cost. This approach has the potential of overcoming inherent problems of the usual approach to audiovisual synthesis, such as asynchrony and incoherence. We also expect the extra benefit of improving the quality of the synthesized acoustic signal, since visual information provides additional relevant distance measures for selection and concatenation.

The results of the preliminary experiments are very promis-

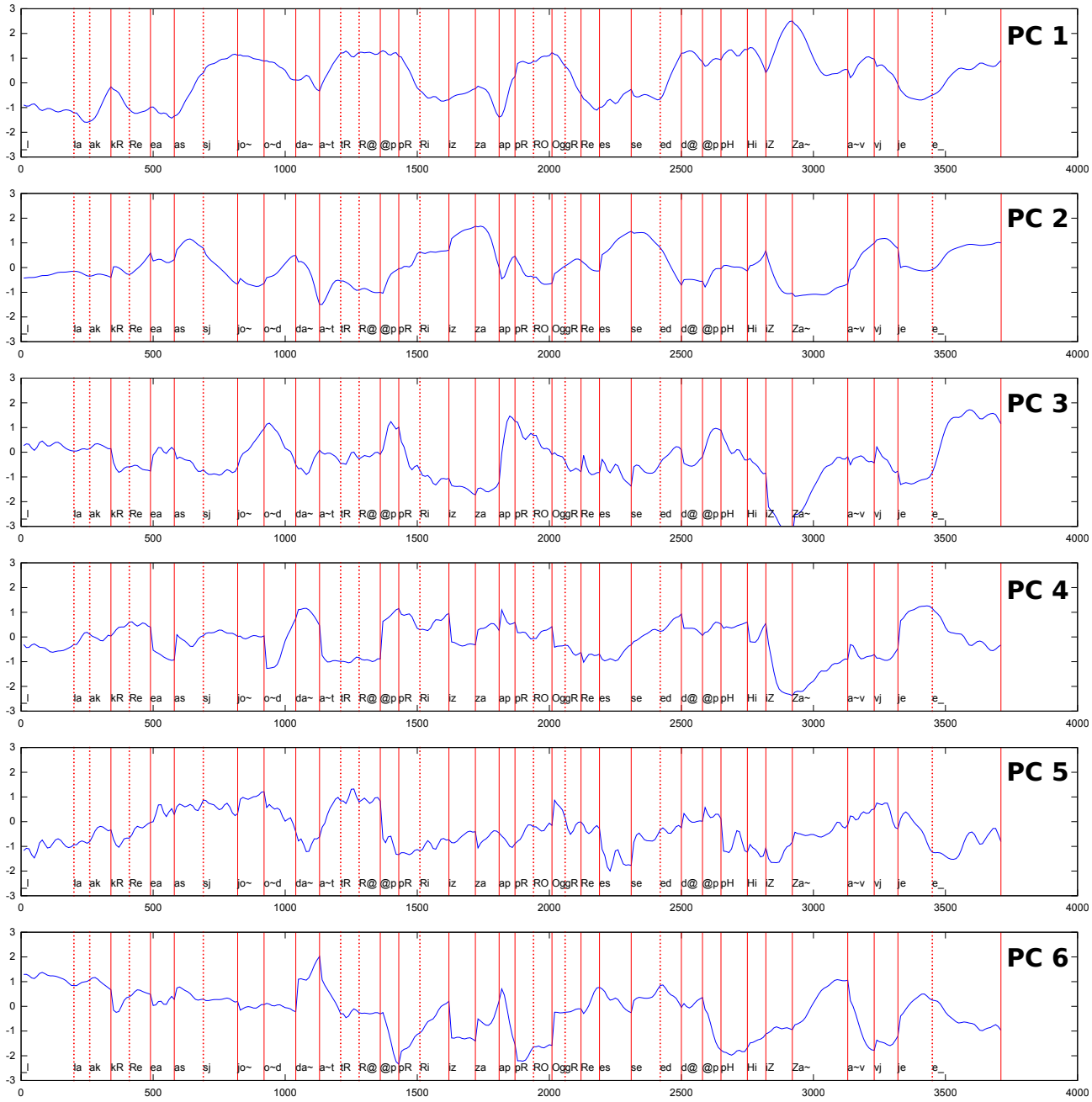

Figure 6: *Concatenated six first visual principal component (in z-scored units) for the test sentence "La création de l'entreprise a progressé depuis Janvier" using a bimodal weighting. Horizontal axes denote time in milliseconds. The boundaries between diphones are marked. Dashed lines indicate that the combination of the two diphones exists consecutively in the corpus and is extracted "as is" from it, solid lines otherwise. SAMPA labels for diphones are shown.*

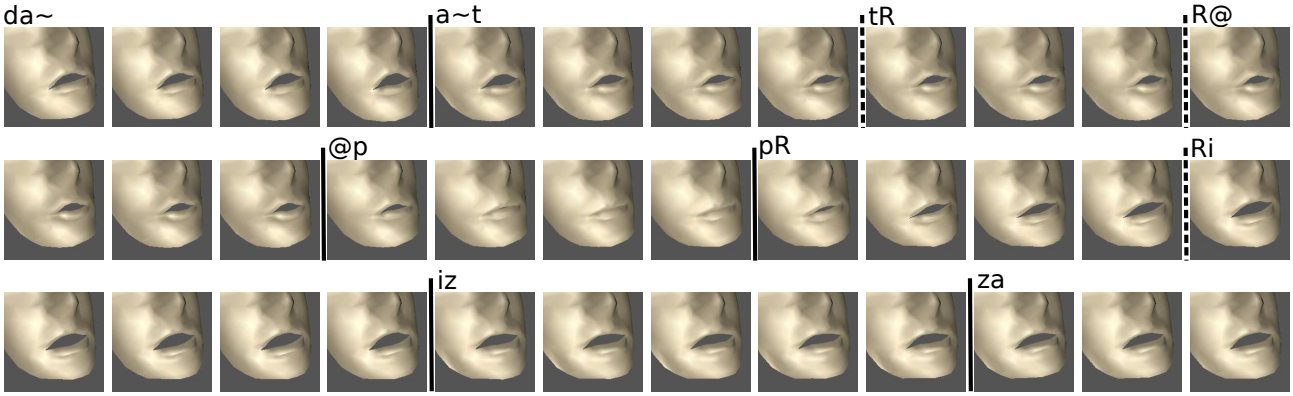

Figure 7: *Sequence of images, derived from synthesized 3D facial information, corresponding to milliseconds 1040 to 1750 of Fig. 5(c). The word depicted in this example is "d'entreprise", and synthesis is done by minimizing the weighted sum of all three costs involved. For the sake of clarity, one image for every 20 ms is shown. Bars mark the boundaries between concatenated diphones (dashed when diphones are consecutive in the corpus).*

ing. In addition, we expect that using a larger corpus (about six times larger than the one used in this study) should improve the quality of the synthesis. We are optimistic that this in itself should increase drastically the quality of the bimodal synthesis, since we have observed such improvement in acoustic-only synthesis experiments with the same up-scaling of corpus size.

In the near future, we will explore and refine the choice of weights in Eqs. 1 and 2. Both of these sets of weights are under questioning, and we have only used preliminary heuristics for setting their values to get the results presented in this paper. It is very likely that a different relative weighting of the three costs (target, acoustic join, and visual join), or a different relative weighting of the contribution of each principal component on the visual cost will improve results.

We also intend to apply appropriate processing of the visual trajectories in the cases where we have some unavoidable mismatch at the boundaries between selected diphones. With the current weights used in the visual cost calculation, it seems that such processing is more relevant for less important principal components.

For the longer term, we are planning a perceptual evaluation of our system, which could also be extended to the evaluation of the integration of the two signals (acoustic and visual). Our results and evaluation can be confronted with other audiovisual speech perception models [16, 17].

6. Acknowledgement

This work was supported by the French National Research Agency (ANR - ViSAC - Project N. ANR-08-JCJC-0080-01).

7. References

[1] J. Barker and F. Berthommier, "Evidence of correlation between acoustic and visual features of speech," in *ICPhS*, San Francisco, USA, 1999.

[2] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1-2, pp. 23–43, 1998.

[3] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[4] B. Le Goff, T. Guiard-Marigny, M. Cohen, and C. Benoit, "Real-time analysis-synthesis and intelligibility of talking faces," in *2nd International Conference on Speech Synthesis*, Newark, NY, USA, 1994.

[5] S. Ouni, M. Cohen, H. Ishak, and D. Massaro, "Visual contribution to speech perception: measuring the intelligibility of animated talking heads," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, p. ID 47891, 2007.

[6] G. Bailly, M. Bérar, F. Elisei, and M. Odisio, "Audiovisual speech synthesis," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 331–346, 2003.

[7] W. Mattheyses, L. Latacz, and W. Verhelst, "On the importance of audiovisual coherence for the perceived quality of synthesized visual speech," *EURASIP Journal on Audio, Speech, and Music Processing*, p. ID 169819, 2009.

[8] A. Hallgren and B. Lyberg, "Visual speech synthesis with concatenative speech," in *AVSP*, Terrigal-Sydney, Australia, 1998.

[9] S. Minnis and A. Breen, "Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis," in *Interspeech*, Bejing, China, 2000.

[10] S. Fagel, "Joint audio-visual units selection the JAVUS speech synthesizer," in *International Conference on Speech and Computer*, St. Petersburg, Russia, 2006.

[11] B. Wrobel-Dautcourt, M. Berger, B. Potard, Y. Laprie, and S. Ouni, "A low-cost stereovision based system for acquisition of visible articulatory data," in *AVSP*, British Columbia, Canada, 2005.

[12] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modelling*, W. Hardcastle and A. Marchal, Eds. Amsterdam: Kluwer Academic Publisher, 1990, pp. 131–149.

[13] V. Colotte and R. Beaufort, "Linguistic features weighting for a Text-To-Speech system without prosody model," in *Interspeech*, Lisbon, Portugal, 2005.

[14] K. Liu and J. Ostermann, "Optimization of an Image-Based Talking Head System," *EURASIP Journal on Audio, Speech, and Music Processing*, p. ID 174192, 2009.

[15] M. Berger, "Realistic face animation from sparse stereo meshes," in *AVSP*, Hilvarenbeek, The Netherlands, 2007.

[16] J. Kim and C. Davis, "Visible speech cues and auditory detection of spoken sentences: an effect of degree of correlation between acoustic and visual properties," in *AVSP*, Scheelsminde, Denmark, 2001.

[17] J. Schwartz, F. Berthommier, and C. Savariaux, "Audio-visual scene analysis: evidence for a "very-early" integration process in audio-visual speech perception," in *Interspeech*, Denver, Colorado, 2002.