Setup for Acoustic-Visual Speech Synthesis by Concatenating Bimodal Units

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

Setup for Acoustic-Visual Speech Synthesis by Concatenating Bimodal Units

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

Université Nancy 2 - LORIA, UMR 7503, BP 239, 54506 Vandœuvre-lès-Nancy, France
{toutiosa,musti,slim,colotte,wrobel,berger}@loria.fr

Abstract

This paper presents preliminary work on building a system able to synthesize concurrently the speech signal and a 3D animation of the speaker's face. This is done by concatenating bimodal diphone units, that is, units that comprise 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 the approach, since both the synthesized speech signal and the face animation are of good quality. Planned improvements and enhancements to the system are outlined.

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

1. Introduction

Speech communication is naturally bimodal. 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].

In the vast majority of recent works, data-driven audiovisual speech synthesis, that is, the generation of face animations together with the corresponding speech acoustics on the basis of recorded data, is still considered as the synchronization of two independent sources: synthesized acoustic speech (or natural speech aligned with text) and the face animation [3]. However, achieving perfect synchronization between these two streams is not straightforward and presents several problems and challenges related to audio-visual intelligibility. Furthermore, it can be the case that the auditory and the visual information originate from different repetitions of the same text, something that may cause problems of perceptual incoherence [4].

To avoid such problems, we propose, within the ViSAC project, to perform synthesis with its acoustic and visible components simultaneously. To this end we consider a bimodal signal as one signal with two channels: acoustic and visual. This bimodality is to be kept together 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 comprise of visual information alongside acoustic information. In our work, 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, which is important toward using the system in the context of applications that involve lip-reading. In addition, we opt for the diphone as the concatenation unit. The advantage of choosing diphones is that the major part of coarticulation phenomena is captured in the middle of the unit and the concatenation is made at the boundaries, which are acoustically more steady. This choice is in accordance with many recent works in concatenative speech synthesis.

The idea of concatenating bimodal units is not entirely new. Earlier studies appeared in [5] and [6]. More recently, two advanced systems based on 2D images were presented in [7] and [4]. These works share several common characteristics with ours. Nevertheless, the combination of features of our system, as presented in this paper, is unique. Our particular ability to acquire and process large amounts of parallel audiovisual data will be very important to keep improving the quality of our results.

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.

We should note that in this paper we present only an initial setup toward our goal of performing bimodal synthesis. We are planning several refinements and enhancements for the immediate future. These will be outlined in the text where appropriate. We give some preliminary results in an attempt to illustrate the potential benefits of our proposed method of combining acoustic and visual constraints in a concatenative system.

2. Data acquisition and modeling

Visual data acquisition was performed simultaneously with acoustic data recording, using a low-cost 3D facial data acquisition infrastructure we developed in the past [8]. 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.

The corpus we acquired for the present work consisted of

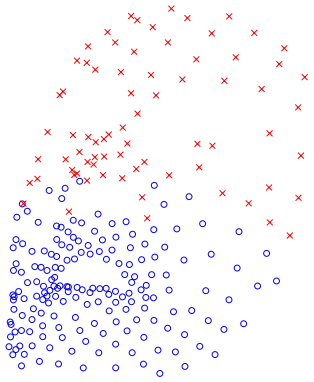

Figure 1: *PCA is applied on the information of 178 out of 252 painted markers (plotted in blue circles). The rest of the markers (plotted in red crosses) are modeled only by their mean value, since they do not reflect explicit speech gestures.*

Table 1: *Percentage of variance explained by the 12 first principal components of facial information. The reference total variance is calculated with respect to the 178 markers retained for PCA (see Fig. 1).*

1: 57.75%	2: 21.93%	3: 6.46%	4: 2.27%
5: 1.55%	6: 1.07%	7: 0.93%	8: 0.56%
9: 0.44%	10: 0.38%	11: 0.33%	12: 0.32%
Total of 12 Components: 93.99%			

the 3D positions of 252 markers with a sampling rate of 188.27 Hz, for 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 speech-derived acoustic parameters. In combination with the sub-sampling, the data were filtered using a low-pass filter with a cutoff frequency of 25 Hz. We found that such 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.

We applied PCA analysis on a subset of markers at the lower part of the face (jaw, lips, and cheeks–see Fig. 1). 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 (see Table 1).

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 lower face visual information 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. A database of diphones was then constructed, including 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 we recorded, as explained previously. The size of this corpus is large enough compared to other works on audiovisual synthesis, but small compared to works on text-to-speech synthesis. We are designing a larger corpus of acoustic and visual data, composed of about 2000 sentences (more than 2 hours of speech). Our goal is that this corpus covers at least all diphones of the French language and contains several representations of these diphones in different contexts. Special care will be taken to account for visual variability alongside acoustic variability.

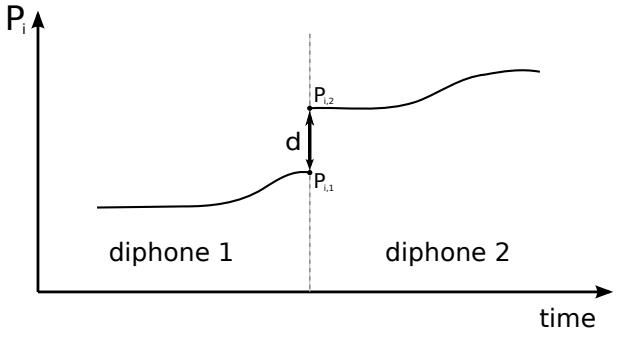

Figure 2: *Illustration of the visual cost calculation. The purpose is to minimize the distance d between the points $P_{i,1}$ and $Pi, 2$ at the boundary of the two concatenated diphones.*

A text to be synthesized is first automatically phonetized and partitioned into diphones. For each diphone, all possible candidates from the database must have the same phonemic label. A special algorithm is available to handle cases when there are no instances of the same diphone in the database. The selection among these candidates is operated by resolution of the lattice of possibilities 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 [9].

Similarly, the visual join cost is defined as the visual distance between the units to be concatenated. This is calculated using the PCA transformed visual information at the boundaries of the units to be concatenated. That is:

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

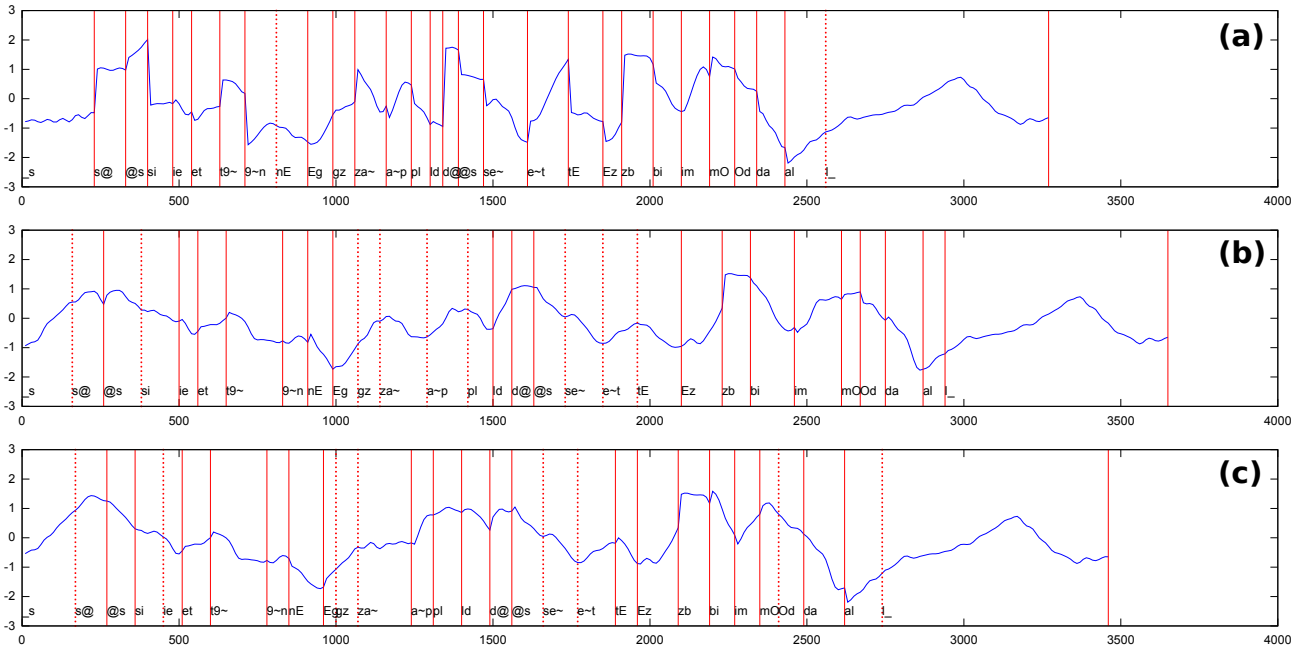

Figure 3: *Concatenated first visual principal component (in z-scored units) for the test sentence "Ceci est un exemple de synthèse bimodale." when: (a) only target and acoustic join costs are minimized; (b) only visual join cost is minimized; (c) a weighted sum of all three costs is minimized. Horizontal 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.*

where $P_{i,1}$ and $Pi, 2$ are the values of the projection on principal component i at the boundary between the two diphones (see Fig. 2). 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 [10], 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 strategy would be helpful to impose continuity upon visual trajectories, at boundaries between diphones, especially for the cases of less important principal components. For the more important components, the trajectories are already acceptably continuous, and we expect that the situation will improve further with a larger corpus. In every case, the high resolution of our data, compared to other works on audiovisual synthesis, will give us additional flexibility to modifying our visual trajectories near diphone boundaries, since we have more samples to manipulate.

4. Results

As a preliminary test for our system, we studied three cases for the weights of Eq. (1). First, we selected $w_{vc} = 0$. That is, we did not apply a visual cost, and the system was driven only by linguistic information and acoustics (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, the target and acoustic costs were set to zero. The system was driven only by visual information (visual-only case).

The third case was truly bimodal. Target and acoustic join 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. 3 we show the synthesized projection on the first principal component for the text string "Ceci est un exemple de synthèse bimodale". In the acoustic-only case (a) there are some obvious discontinuities in the visual trajectory. These discontinuities result 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 sufficiently continuous. This is true for the first component, however, as we move to less important components (not shown) discontinuities start to appear gradually. This is because the weights (the eigenvalues–see also Table 1) we selected for Eq. (2) put a lot of emphasis on the first few components. Nevertheless, the more important a principal component, the more sensitive the talking head animation is to discontinuities in the corresponding projection. We should also notice that both visual-only and bimodal cases seem to favor continuity as more pairs of diphones appearing consecutively in the database (presented with dashed lines in Fig. 3) were selected for synthesis than in the acoustic-only case. This fact may indicate a weakness in our acoustic-only synthesis setup which is however at large corrected with the inclusion of visual constraints.

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 still intelligible, has several problems regarding duration of diphones, intonation and some audible discontinuities at boundaries between diphones. On the other hand, the waveform syn-

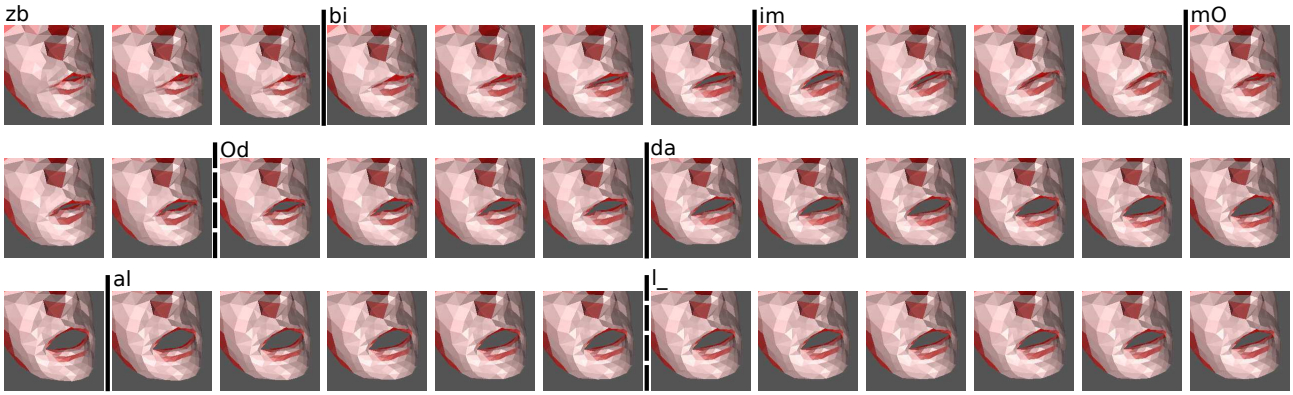

Figure 4: *Sequence of images, derived from synthesized 3D facial information, corresponding to milliseconds 2140 to 2840 of Fig. 3(c). The word depicted in this example is "bimodale", and synthesis is done by minimizing the weighted sum of all three costs involved. For sake of clarity, one image for every 20 ms is shown. Bars mark the boundaries between concatenated diphones (dashed when diphones are consecutive in the corpus).*

thesized at the bimodal case is at large free of such problems.

Broadly, using in conjunction acoustic/linguistic and visual constraints combines benefits from using in isolation either acoustic/linguistic or visual constraints. Fig. 4 shows an animated sequence using all constraints for a part of the sentence shown in Fig. 3. 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 in the duration of our project, something that will add more realism to our final results [11].

5. Conclusion

We have presented an initial setup and preliminary experiments toward building a system able to perform talking head synthesis with its acoustic and visible components simultaneously, on the basis of concatenating bimodal diphones, that is, units that comprise both acoustic and visual information. 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.

We are planning several actions for the near future to improve upon this initial system. The most important is the design and exploitation of a corpus about six times larger than the one we are currently using. We are optimistic that this alone will improve vastly the quality of our results, since we have observed such improvement in acoustic-only synthesis experiments with the same up-scaling of corpus size.

We also need to explore and refine more the choice of weights in Eqs. (1) and (2). Both of these sets of weights are under questioning, and we have only used initial heuristics for setting their values to get the results we presented in this paper. Perhaps 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 lead to improved results.

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

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] 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.

[4] 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*, 2009.

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

[6] 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.

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

[8] 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.

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

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

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