Introducing Visual Target Cost within an Acoustic-Visual Unit-Selection Speech Synthesizer

Utpala Musti, Vincent Colotte, Asterios Toutios, Slim Ouni

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

Abstract

In this paper, we present a method to take into account visual information during the selection process in an acoustic-visual synthesizer. The acoustic-visual speech synthesizer is based on the selection and concatenation of synchronous bimodal diphone units i.e., speech signal and 3D facial movements of the speaker's face. The visual speech information is acquired using a stereovision technique. Unit selection for synthesis is based on the classical target cost consisting of linguistic and phonological features. We compare several methods to take into account the visual articulatory context in the target cost. We present an objective evaluation of the synthesis results based on correlation of the actual visual speech trajectory and synthesized visual speech trajectory.

Index Terms: speech synthesis, unit selection, target costs.

1. Introduction

Face-to-face communication is more effective than situations involving just the voice. The main reason is that the face improves intelligibility [1], particularly for hard of hearing and language challenged population [2]. The framework of audio-visual speech synthesis allows the development and study of talking heads. In this context, data-driven synthesis is gaining popularity in audiovisual speech synthesis. The advantage is that the natural facial gestures are captured which may help in preserving this naturalness in the final result. However, the focus is usually made on visual animation, and the final synthesis is a synchronization of this animation with audio obtained by a text-to-speech system. Thus the two components are independent. In our paradigm, we keep the two components together during the whole process, and thus we consider them as one bimodal signal with two facets, acoustic and visual. This has the advantage of keeping the coherency between both components and avoids any perceptual ambiguity. Our acoustic-visual text-to-speech synthesis (TTS) approach follows unit selection acoustic TTS synthesis technique. In this scheme, good quality acoustic speech is synthesized based on the selection and concatenation of units from a large database [3]. To perform bimodal synthesis we have extended this idea to concatenate its acoustic and visual components simultaneously. The first works dealing with the concatenation of bimodal units are those described in [4] and [5]. Recently, two systems based on 2D images were presented in [6] and [7]. The particularity of our system [8] is that it is based on 3D meshes where the lower part of the face, related to speech gestures, is using an important number of markers.

The significant step in the unit selection technique is the appropriate selection of target units satisfying a given target specification. The features classically used for this purpose include linguistic, phonetic and prosodic context. This set of features is language-dependent. The phonetic context also gets affected significantly by speakers' articulatory preferences and idiosyncrasies. This information is usually based on characterization defined by phoneticians and found in literature. However, due to the usage of a recorded audio-visual corpus, in case the speaker has a peculiar articulation, it might be visually perceived during synthesis and the final result might present some incoherence. Actually, in our case, quantitative information regarding articulation is available and directly measurable from the recorded audiovisual corpus. For this reason, we propose in this work to modify the phonetic context information based on how the recorded speaker is really articulating phonemes. The main goal is to see whether this improves the performance of synthesis.

Measuring the improvement is not an easy task and developing some evaluation technique is needed. Such a technique should make it possible to assess the synthesis results of any relevant modifications to the target cost or any other modifications in other components like visual and acoustic join costs and their weights. The perceptual evaluation techniques are time consuming and laborious and do not allow for quick evaluation of any particular strategy. Perceptual evaluation should be considered for a global and final evaluation of the system, and not during the development process. An objective evaluation using a metric to evaluate the synthesis results reflecting the perceptual similarity to the best possible extent is a better way to approach the evaluation problem. One of the common objective evaluation measures used for the acoustic concatenative speech synthesis for the trajectory comparison of real and synthesized sentences is cepstral distance between waveforms [3]. In this paper, we focus on visual domain objective evaluation by determining the Pearson correlation coefficient and root mean square error (RMSE) between the synthesized and real trajectories of the visual features. A similar objective evaluation has been performed in [9] though the overall system and the approach were completely different. In their system, an HMM trained for lip movements with the audio-visual database is used to generate trajectory which is then used as a guide for the selection of optimal mouth images.

In the following sections we first present an overview of our setup for acoustic-visual speech synthesis. Then we present the ideas related to the modification of the phonetic context definition based on the corpus characteristics, followed by the visual domain objective evaluation.

2. Overview of Bimodal Synthesizer

2.1. Data acquisition and modeling

The bimodal speech corpus consists of visual speech and acoustic speech recorded simultaneously. The visual speech information consists of 252 3D point trajectories acquired using stereovision technique sampled at 188.27 Hz. It is acquired simultaneously with acoustic speech recorded at 16kHz with 16-bit precision. The corpus consists of 319 medium-sized French sentences, covering about 25 minutes of speech. PCA was applied on a subset of the 3D point data consisting of lower part of the face (jaw, lips and cheeks) and first 12 principal components are retained which account for 94% of the data variance.

2.2. Bimodal unit selection and concatenation

For the synthesis of bimodal speech from text, first the text is automatically phonetized and partitioned into diphones. The diphone units are looked up in a diphone database built on the corpus. The diphone database is built by phonetizing and partitioning into diphones after linguistic analysis. For each diphone required for synthesis, all possible candidates from the database having the same phonetic label are looked up. A special algorithm is available to handle cases where there are no instances of the same diphone in the database.

The target specification used for look up and ranking the diphone candidates consists of linguistic and phonetic features which specify the phonemes being looked up and their linguistic and phonetic content and context which affect their realization. The target specification is compared with the candidate description for the calculation of the target cost. In addition to linguistic features (word position, syllable position etc) the phonetic context is taken into account as binary values in the target cost. Each of the contextual phonemes is classified as belonging to an articulatory category of phonemes (protruded, spread, etc). For instance, the phoneme /u/ belongs to the set of protruded phonemes for French. This kind of discrete classification is based on classical phonetic knowledge [11]. Thus, a candidate with the same articulatory context like the target phoneme will not be penalized in the target cost i.e. that target cost component would be zero, else one. In our system, the target cost of a diphone candidate is just the summation of target costs of the two phonemes composing this diphone. The target cost of each of the phonemes is a weighted summation of the difference between the features of the candidate phoneme and the features of the target phoneme. The selection among the set of selected 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 (AJC), and the visual join cost (VJC), that is

$$C = w_{tc}TC + w_{ajc}AJC + w_{vjc}VJC \qquad (1)$$

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

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 target feature weighting and calculation of the target and join costs, see [12]. 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.

The selected diphone sequence is concatenated in both modalities. Further details about our bimodal speech synthesizer can be found in [8].

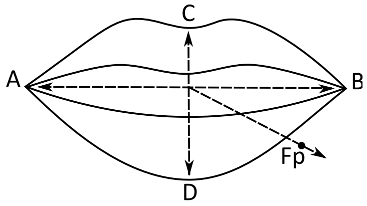

Figure 1: *Articulatory features. Mainly labial using 4 markers on the face: A, B, C and D. Lip opening and lip spread are given by the distances $\|\vec{CD}\|$ and $\|\vec{AB}\|$. Lip protrusion is given by the displacement of O, the barycenter of the four points (A, B, C, D) along the normal vector ($\vec{OFp}$) to the plane formed by vectors $\vec{AB}$ and $\vec{CD}$.*

3. Phonetic Context Adaptation

The phonetic context of any particular phoneme influences the articulation significantly. This is well known as coarticulation. The degree by which a phoneme influences its surrounding phonemes or is influenced by them varies [13]. The established phonetic knowledge regarding coarticulation holds almost all the time [10, 11] . However, as mentioned above, if for some reason the speaker has a peculiar articulation strategy, this might be visually perceived during synthesis and the final result might present some incoherence. It seems natural, as we are dealing with the acoustic-visual synthesis, to take into account the recorded visual information regarding articulation that is available and directly measurable from the recorded audiovisual corpus. For instance, the articulatory features like lip opening, lip protrusion and lip spreading are indeed considered as visual features.

In the current scope, we propose to take into account the articulatory strategy of the recorded speaker to determine more precisely the category of a given phoneme. The modification of the phonetic context should modify the visual target cost, which is a part of the target cost (TC). The visual target cost of a phoneme (left or right phoneme of a diphone) is calculated by summing the visual feature differences of the left and the right contextual phonemes.

3.1. Phonetic category modification

The purpose of the following experiment was to change the current characteristics of some phonemes which were based on phonetic knowledge. The changes modified the target and candidate description for the target cost to better take into account their main characteristics as observed in the audio-visual corpus. The expectation was that their synthesized visual speech component would be more similar to the real visual speech after the changes. The phonetic feature cost describing the context of a candidate has a binary value: 0 if phonetic feature is equal to that required by the target, otherwise 1.

In this work, we considered the lip shape and the place of articulation as the set of target features that affect the visual coarticulation. For a given phoneme, the lip shape takes only 3 values: {protruded, spread and none}. We have tried to change this classification for some phonemes based on the analysis of the visual data of the corpus. We performed some statistical analysis of the articulatory features in comparison with the phoneme categories. More specifically, we re-examined the classification of phonemes initially categorized as ‘protruded’, ‘spread’ and

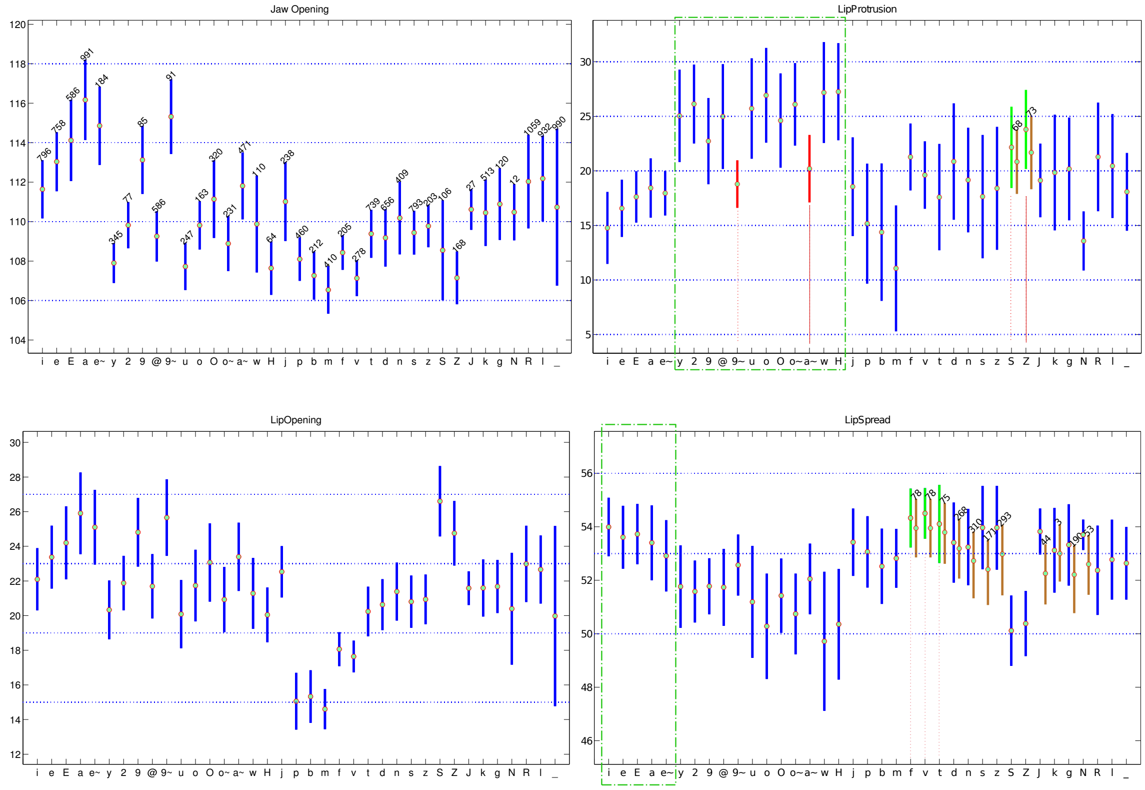

Figure 2: *Articulatory feature statistics: Each segment represents a phoneme, centered at the mean and its length is twice the standard deviation. The number of occurrence of each phoneme is presented. The phonemes of interest are framed : (1) the 'protruded' phonemes* {*y, ø, œ, ə, õe, u, o, õ, ɔ , ã, w, ɥ* }*; (2) the 'spread' phonemes* { *i, e, a, ɛ, ẽ*}*. All the other phonemes are labeled 'none'. The segments plotted in red and green correspond to the phonemes that their category was modified. The brown segments are those where statistics were recalculated without protruded context in LipProtrusion graph, and without spread context in LipSpread context.*

'none'. As shown in Figure 1, the set of articulatory features included lip protrusion, lip opening, lip spreading and jaw opening [14]. The statistics were calculated by considering the articulatory feature vectors at the center of the phoneme articulation (place of concatenation in the visual and acoustic domain).

Typically, the set of phonemes which included { i, e, a, ɛ, ẽ} was categorized as 'spread' phonemes and the following set of phonemes {y, ø, œ, ə, õe, u, o, õ, ɔ , ã, w, ɥ } was classified as 'protruded'. All other phonemes were considered as 'none' based on the shape of the lips.

The statistics of the phonetic articulatory features are shown in figure 2. We considered the mean, the variance and the number of occurrence of each phoneme. This figure shows that overall the categories are respected. Nevertheless, we can observe that some phonemes need to be reconsidered. For this purpose and to be more accurate, the coarticulation effects of the surrounding phonemes should be removed. In fact, if one of the neighboring phonemes is protruded, for instance, it is very likely that the surrounded phoneme will be protruded too, even if it is not its main articulatory characteristic, because of coarticulation. Therefore, for phonemes whose visual articulation seemed to be different from their initial classification, their articulatory feature statistics were recalculated by considering a subset of phoneme instances in the corpus. For example, the phoneme /f/ seemed to be 'spread' unlike its classical phonetic classification of 'none'. Thus, only its occurrences in the corpus without spread phonemes in its neighborhood were taken into account. Its articulatory feature statistics were recalculated to confirm its effective visual articulation. The following set of phonemes were considered for recalculation to check if their effective articulation is 'spread': {f, v, t, d, n, s, z, ɲ, k, g, ŋ }. For the two phonemes {ʃ and ʒ}, the articulatory feature statistics without rounding context was recalculated. Initially, the sets of phonemes {f, v, t}, {ʃ, ʒ} and {ã, õe} were considered as 'none', 'none' and 'protruded' respectively. However, based on the statistics and the observation of the data, we found out that the strategy of our speaker is quite different from this definition. For this reason, we modified the articulatory target features for these sets phonemes to 'spread', 'protruded' and 'none' respectively.

In section 4, we present an evaluation where we compared the synthesis using the initial articulatory description (IPD) and the changed phonetic description (CPD).

3.2. Continuous visual target cost

From the statistics of the articulatory features explained in the previous section we reclassified phonetic characteristics into distinct categories. The goal was to adapt the classification to the real ones based on the corpus used. But one can observe that it is not easy to take a discrete distinct decision from these statistical values. So the visual target cost component has to be formulated as a real value in the range $[0, 1]$ rather than binary value. In this way, the articulatory characteristics should be considered as continuous. So the visual target cost component has to be formulated as a real value in the range $[0, 1]$ unlike binary value. For calculating the continuous target cost we used the articulatory feature statistics calculated as explained in section 3.1. In the following subsections we first explain the earlier work by Mattheyses et al. [15] towards a continuous visual target using a phoneme difference matrix and then our method of calculating the continuous visual target cost. In [15], the authors used shape and texture parameters extracted by applying Active Appearance Models on 2D facial images of speech animation. We tried to apply the same logic for the calculation of the continuous target cost using articulatory features.

The articulatory feature statistics are represented by μ_{ij} and σ_{ij} to represent the means and variances for all the phonemes (index i) and the various articulatory features (index j).

3.2.1. Visual target cost based on phoneme difference matrix

In [15], the calculation of visual target cost was as follows: Two phonemes were considered similar in terms of visual representation if their mean representations were alike and, in addition, if these mean representations were sufficiently reliable (i.e. if small summed variations were measured for these phonemes). Two matrices were calculated, which express for each phoneme pair (p, q); the difference between their mean representations D^{μ}_{pq} and the sum of the variances of their visual representation D^{σ}_{pq}, respectively:

$$D^{\mu}_{pq} = \sqrt{\Sigma_j(\mu_{pj} - \mu_{qj})^2}$$

$$D^{\sigma}_{pq} = \Sigma_j \sigma_{pj} + \Sigma_j \sigma_{qj}$$

Scaling both matrices between zero and one gave $D^{\mu'}_{pq}$ and $D^{\sigma'}_{pq}$, after which the final difference matrix was calculated:

$$D_{pq} = 2D^{\mu'}_{pq} + D^{\sigma'}_{pq}$$

Matrix D_{pq} is used to calculate the visual target cost during selection.

3.2.2. Visual target cost based on contextual significance

In the previous method, the point of emphasis was centered on the contextual phonemes without taking into account the nature of the main target phoneme. For each phoneme, the feature with least variance is the one which gets least modified due to coarticulation and the features with higher variance get affected more due to coarticulation. Thus, obtaining similar context is important for features which get more influenced due to coarticulation. We applied this principle for the calculation of phoneme difference $D_{pq}(i)$ as function of the target phoneme i which is looked up in the corpus; where, p is the contextual phoneme (left or right) of i in the target utterance and q is the contextual phoneme of the candidate for i. The difference of the mean of the contextual phoneme was weighted by the variance of the target phoneme:

$$D_{pq}(i) = \Sigma_j w_{ij} |\mu_{pj} - \mu_{qj}| \qquad (2)$$

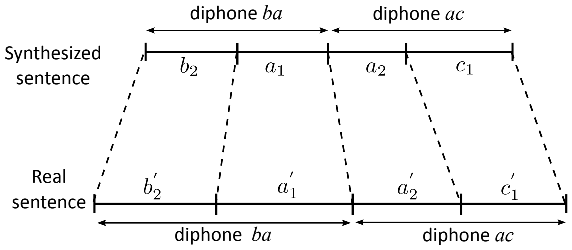

Figure 3: *Adjusting diphone lengths. Each of the corresponding half-phones which are part of the diphones in the synthesized and real sentences are re-sampled through linear interpolation to make the number of visual samples equal.*

$$w_{ij} = (\frac{\sigma_{ij}}{\Sigma_j \sigma_{ij}})$$

$D_{pq}(i)$ was scaled between zero and one. This gives the distance between contextual phonemes as a function of the phoneme for which the proximate context is being looked up during the selection process for the calculation of visual target cost.

The weight w_{ij} gives the relative importance of the component j with respect to the others. More the variance σ_{ij} is, higher the weight on the contextual difference for the component j is. Thus, w_{ij} reflects the fact that context has important impact on these components with higher variance.

4. Objective Evaluation of Synthesis and Results

4.1. Objective Evaluation

For the purpose of evaluating the synthesis results, we used a method based on leave-one-out cross-validation technique, where a single sentence from the corpus (the set of 319 sentences) was used to evaluate the synthesis (comparing the synthesized sentence with this original one) and the remaining sentences of the corpus as source for the database of diphones for the acoustic-visual synthesizer. This was repeated in a way that each sentence was used once as a test sentence. The advantage of this method is that it avoids building a specific test corpus for evaluation. However, we reduce marginally the choice of selection, by removing some diphones of the sentence discarded from the original corpus.

After synthesizing a given sentence all the half-phones (the two half-phones of a diphone) of the synthesized sentence and the actual sentence were re-sampled individually to make the number of visual samples equal in both the real and synthesized sentences (see figure 3). This was done using a simple linear interpolation of the 12 PCA coefficients. After this, the Pearson correlation coefficient between 12 PCA coefficients of all the synthesized sentences and the real sentences actually present in the corpus was determined. Similarly, the Pearson correlation coefficient between 4 articulatory parameters was also determined. The root mean square error between articulatory feature trajectories of the synthesized and the real sentences present in the corpus was determined.

If x_d and y_d are the sequences of the d^{th} PCA coefficient of a real and synthesized sentence having n samples, the correlation coefficient is calculated as follows:

$$r_{x_d y_d} = \frac{n\Sigma x_d(i)y_d(i) - \Sigma x_d(i)\Sigma y_d(i)}{\sqrt{n\Sigma x_d(i)^2 - (\Sigma x_d(i))^2}\sqrt{n\Sigma y_d(i)^2 - (\Sigma y_d(i))^2}} \quad (3)$$

Though it is almost impossible to have a perfect correlation between the real and synthesized sentence, it seems to be a reasonable assumption that the trajectories for two diphones selected with similar phonetic context and linguistic description would be significantly correlated or similar. For now, we proceed with visual domain objective evaluation of the speech information alone, assuming that the visual speech animation would be strongly correlated with the underlying acoustic speech. Alongside, we are working on developing other metrics where both acoustic and visual components are considered simultaneously. An example of the trajectories of the first 3 principal components of a synthesized sentence and the corresponding real sentence is shown in figure 4.

4.2. Evaluation Results

Based on the above explained objective evaluation technique the performance of the various visual target cost techniques were determined. The target cost techniques with the binary visual target cost components (IPD and CPD) performed comparable to each other ($r_{x_d y_d} = 0.813$ for PC 1). Similarly, the continuous visual target cost components phoneme difference matrix approach (PDM) and phoneme difference based on contextual significance (PDCS) performed comparable to each other ($r_{x_d y_d} = 0.816$ for PC 1). The continuous visual target costs gave marginally better results consistently compared to the binary visual target cost approaches even when different weights for the visual target cost component were used. This is also apparent when observing the performance with respect to articulatory features. In fact, the correlation for the first two methods IPD and CPD is 0.70 and it increases up to 0.72 for the PDM and PDCS for jaw opening (see table 1). Table 2 shows the RMSE between real and synthetic trajectories for the articulatory features. The RMSE is almost the same for the 4 methods. We should notice that each of the examined methods affects the ranking of the selected candidates though it is not that obvious that there are differences between them. We should emphasize that the significance of this examined visual target cost component in the overall target cost is 1%, as we have a large set of features. Therefore this can explain this marginal variation in the performance.

Hence, the possibility that a continuous target cost component better represents the differences between phonemes while optimizing the synthesis performance for particular corpus than discrete binary target cost components has to be contemplated. Given the limited generalizing power, for a corpus of small size and without a very well balanced diphone coverage in the corpus, the target cost consisting of phonetic context modeling based on classical knowledge can be considered sufficient. One should observe that the objective evaluation used in this work is purely visual. It will be more appropriate if we consider a combination of visual and acoustic metrics to be used for such evaluation. This is actually our current focus, however finding the balance between the acoustic component and the visual one is still not easy to reach and is under careful consideration.

Examining the results of the objective evaluation presented in this paper of our acoustic-visual synthesis, shows that they are quite good. The overall correlation is quite high. In addition, the RMSE is very low and acceptable. In fact, the Jaw opening RMSE is around $2mm$, lip opening ($2.7mm$), lip spreading ($1.38mm$) and lip protrusion is $4mm$. This is a good indication that our synthesis method provides similar trajectories to those of real sentences. This is quite interesting, as we know that the purpose of synthesis is not to generate the exact speaker articulation (unlike acoustic-to-articulatory inversion). Thus, it seems that our acoustic-visual synthesis, based on the main idea of considering the signal as one bimodal signal, was able to capture finely the articulatory strategy of our speaker. This can be clearly seen in Figure 4.

Table 1: Correlation coefficients between the real and synthesized trajectories of first 3 principal component coefficients and the four articulatory features by various target cost strategies. IPD: initial phoneme description, CPD: changed phoneme description, PDM: phoneme difference matrix, PDCS: phoneme difference based on contextual significance. The articulatory features: JO (jaw opening), LP (lip protrusion), LO (lip opening) and LS (lip spreading). The first four principal components account for about 58%, 24% and 7% respectively.

PC	IPD	CPD	PDM	PDCS
1	0.813	0.813	0.816	0.816
2	0.715	0.715	0.719	0.720
3	0.726	0.725	0.729	0.729
JO	0.708	0.708	0.728	0.728
LP	0.694	0.693	0.698	0.698
LO	0.671	0.670	0.689	0.689
LS	0.636	0.636	0.640	0.640

Table 2: Root Mean Square Error (RMSE) in millimeters between the real and synthesized trajectories of the four articulatory features (same notations as table 1).

AF	IPD	CPD	PDM	PDCS
JO	2.11	2.11	2.06	2.06
LP	4.04	4.04	4.02	4.02
LO	2.70	2.70	2.63	2.63
LS	1.38	1.38	1.37	1.37

5. Conclusion

In this paper, the inclusion of the articulation strategy of the recorded speaker within the unit selection step was examined. The evaluation results show that there is scope for improving the synthesis performance by optimizing the target cost function for the underlying corpus being used for unit selection. The improvement is marginal for now. However, we believe that it is promising and worth more investigation. The visual target cost of the phonetic context is a small portion of bigger set of features, where each feature has its own weight. In addition, the current corpus is still a small corpus. Thus, tuning weights and evaluating the different strategies on a larger corpus might provide more insights. Currently, we are working on larger audiovisual corpus that we will use in the near future. The simple objective evaluation, visual only though, showed that our acoustic-visual synthesis method has the capability to capture the speakers articulatory strategy. It is very likely that considering the acoustic-visual signal as bimodal is very interesting approach.

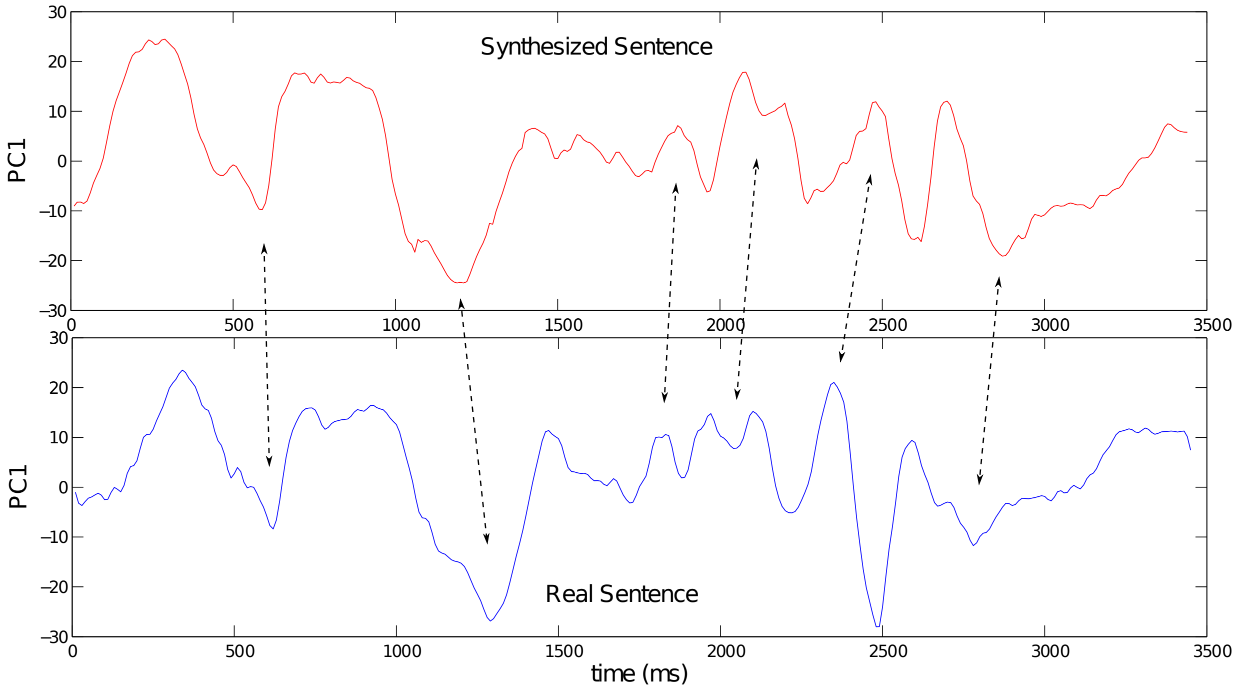

Figure 4: *Synthetic and Real trajectories for the first principal component for the sentence “Sur ces mots, elle sortit vivement de la pièce.” with the following phoneme sequence “sil* s y ʁ s e m o *sil* ɛ l s ɔ ʁ t i v i v ə m ã d ə l a p j ɛ s *sil”. The Pearson correlation for the first principal component was* 0.89.

6. Acknowledgement

Our work was supported by the French National Research Agency (ANR - ViSAC - Project N. ANR-08-JCJC-0080-01). The 3D data used in this project were acquired by Brigitte Wrobel-Dautcourt and Marie-Odile Berger (Magrit group).

7. References

[1] 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.

[2] D. W. Massaro, “Embodied agents in language learning for children with language challenges,” *Computers Helping People with Special Needs*, pp. 809–816, 2006.

[3] A. Hunt and A. Black., “Unit selection in a concatenative speech synthesis system using a large speech database,” *ICASSP-96*, vol. 1, pp. 373–376, 1996.

[4] A. Hallgren and B. Lyberg, “Visual speech synthesis with concatenative speech,” *AVSP*, 1998.

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

[6] S. Fagel, “Joint audio-visual units selection the JAVUS speech synthesizer,” *International Conference on Speech and Computer*, 2006.

[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*, 2009.

[8] A. Toutios, U. Musti, S. Ouni, V. Colotte, B. Wrobel-Dautcourt, and M.-O. Berger, “Setup for acoustic-visual speech synthesis by concatenating bimodal units,” *AVSP*, 2010.

[9] W. Lijuan, Q. Xiaojun, H. Wei, and K. S. Frank, “Photo-real lips synthesis with trajectory-guided sample selection,” *Speech Synthesis Workshop*, 2010.

[10] P. Ladefoged, *A Course in Phonetics.*, 2nd ed. New York: Harcourt Brace Jovanovich, 1982.

[11] P. Ladefoged and I. Maddieson, *The Sounds of the World’s Languages*. Malden: Wiley-Blackwell, 1995.

[12] V. Colotte and R. Beaufort, “Linguistic features weighting for a Text-To-Speech system without prosody model,” *Interspeech*, 2005.

[13] A. Löfqvist, “Speech as audible gestures,” in *Speech Production and Speech Modelling*, W. Hardcastle and A. Marchal, Eds. Dordrecht: Kluwer Academic Publishers, 1990, pp. 289–322.

[14] V. Robert, B. Wrobel-Dautcourt, Y. Laprie, and A. Bonneau, “Inter speaker variability of labial coarticulation with the view of developing a formal coarticulation model for french,” *In AVSP*, pp. 65–70, 2005.

[15] W. Mattheyses, L. Latacz, and W. Verhelst, “Optimized photorealistic audiovisual speech synthesis using active appearance modeling,” *In AVSP-2010*, pp. 145–150, 2010.