Weight Optimization for Bimodal Unit-Selection Talking Head Synthesis

Asterios Toutios, Utpala Musti, Slim Ouni, Vincent Colotte

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

Abstract

This paper addresses talking head synthesis based on the concatenation of units comprising of both acoustic and visual information. Selection of appropriate diphone units to synthesize a given text string is based on the minimization of a weighted linear combination of four costs that reflect linguistic, acoustic, and visual considerations. We present initial work toward a method to determine automatically the weights applied to each cost, using a series of metrics that assess quantitatively the performance of synthesis.

Index Terms: talking head, audiovisual speech synthesis, selection, optimization

1. Introduction

Complementing synthesized speech acoustics with an animation of the speaker's face offers to the listener improved intelligibility in noisy environments, better comprehension of the speech signal, and an increased feeling of positivity and confidence [1, 2]. Possible applications of such talking heads include, among others, the development of virtual assistants, and of language training systems for the hearing impaired [3].

The usual approach to audiovisual speech synthesis considers the synchronization of two independent sources: synthesized acoustic speech (or natural speech aligned with text) and the face animation (for recent examples see [4, 5]). This approach may present perceptual incoherence, since it can be that the auditory and visual information originate from different utterances of the same spoken text [6]. We are carrying research toward a system that overcomes this problem by using an alternative approach [7]: audiovisual synthesis is performed with its acoustic and visible components simultaneously, by considering a bimodal signal comprising of an acoustic and a visual channel. The setup is similar to typical acoustic-only unit-selection synthesis [8], using the diphone as the concatenation unit. However in our case units include both acoustic and visual information.

Selection of appropriate units to synthesize a given text string involves the minimization, through Viterbi search, of a weighted linear combination of four costs that reflect linguistic, acoustic, and visual considerations. The chosen weighting of these costs affects the quality of the synthesized talking head. In our previous work [7], as in other works that adopt similar strategies to talking head synthesis (e.g. [6, 9]), the costs were chosen manually.

In acoustic-only speech synthesis, automatic fine-tuning of such weights is a difficult problem and still an active research field [10, 11, 12] where proposed methods try to provide an objective metric of perceptual cues, to later minimize it over the weights. In audiovisual synthesis, the problem is even more difficult since there is an additional modality (visual) of different nature compared to acoustics.

Usually, proper evaluation of audiovisual synthesis is done via perceptual experiments, but this is not suitable for automatic fine-tuning of the parameters involved in synthesis. What is needed to this end is to investigate objective, quantitative, metrics to assess the synthesis results. The lack of such means of assessment partly explains the need for manual fine-tuning, which is expensive, subjective, and error-prone.

In this paper we present our work toward such objective evaluation and fine-tuning of parameters for talking head synthesis. First, we briefly present our system and the main improvements compared to what we have presented previously [7]. These improvements are the introduction of a derivative visual cost and the application of a special algorithm at the concatenation step to improve the visual joins between diphones. We then introduce a series of metrics to assess the quality of synthesis, in both the acoustic and visual domains. Finally, we merge these metrics into a single one and use a nonlinear optimization technique to minimize it, by adjusting the parameters involved in synthesis, over a set of test utterances. Thus, we arrive at an optimized set of parameter values, which is considerably different with respect to previous heuristic guesses of ours.

2. Acoustic-Visual Synthesis System

Our corpus consisted of the 3D positions of 252 markers painted on the face of the speaker and the concurrently recorded speech signal, for 319 medium-sized French sentences, covering about 25 minutes of speech, uttered by a native male speaker. The positions of the markers were captured using a low-cost 3D facial data acquisition infrastructure [13], with a sampling rate of 188.27 Hz. Acoustics were recorded at 16 kHz with 16-bit precision. Visual data were sub-sampled to 100 Hz, for easier labelling and alignment with speech-derived acoustic parameters. Principal component analysis was applied on the positions of 178 markers at the lower part of the face (jaw, lips, and cheeks) and 12 principal components (PC) were retained. Thus, visual information was reduced to a set of 12 trajectories for each utterance. The corpus was phonetized, analyzed linguistically, and partitioned into diphones. A database was then constructed, including information on position, duration, acoustic, visual (PC trajectories and derivatives) and linguistic parameters for each diphone.

At execution time, 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

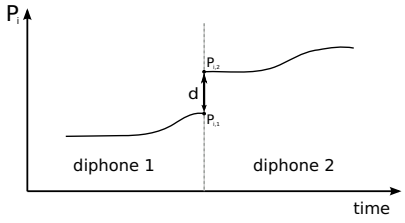

Figure 1: *Illustration of the visual cost calculation. The purpose is to minimize the distance d between the points $P_{i,1}$ and $P_{i,2}$ at the boundary of the two concatenated diphones.*

four costs, i.e.

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

where TC is the target cost, JC is the acoustic join cost, VC is the visual join cost, and DVC is the derivative visual join cost. TC 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 [14]. JC 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. VC is calculated using the values of the PC trajectories at the boundaries of the units to be concatenated, i.e.

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

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. 1). The weights w_i should reflect the relative importance of the components, and we choose them to be proportional to the eigenvalues of PCA analysis, in accordance with [15]. DVC is calculated in the same manner as VC, only using the derivatives of the PC trajectories instead of the trajectories themselves. Derivatives were calculated using a five-point stencil approximation.

In the acoustic domain, the selected diphone sequence is concatenated using a well-studied technique, where pitch values are used to improve the join of diphones. In the visual domain, we apply an adaptive local smoothing around joins which present discontinuities. If the first (Δ) or second ($\Delta\Delta$) derivatives at a given sample of a synthesized visual trajectory lie out of the range defined by ± 3 standard deviations (measured across the whole corpus) then this sample is judged as problematic. We traverse a visual trajectory x_i, and check Δ and $\Delta\Delta$ at each sample i. If one of them is out of the desired range, we replace samples x_{i-k} to x_{i+k} by their 3-point averaged counterparts, using incremental values for k, until Δ and $\Delta\Delta$ at sample i are within the desired range.

3. Metrics for Assessment

The values over a synthesized utterance of the costs involved in Eq. (1) could be used as indicators of the quality of the synthesis results. For instance a smaller JC value indicates a better overall acoustic join. However, it would be problematic to use these in our attempt to fine-tune the weights involved in Eq. (1) for two reasons: first, using the costs in order to determine the weights applied on the same costs, would essentially lead to a recursive loop; second, our aim is to prevent some specific problems related to the quality of the synthesized utterances rather than minimizing overall costs.

As exemplified in Fig. 1, for two selected consecutive diphones, there is a potential difference between the rightmost value of the visual trajectory for the left diphone and the leftmost value for the right diphone. If this difference is large enough, the face animation may be jerky. Processing at the concatenation step alleviates such problems, however a general rule of thumb for unit-selection synthesis is that the selection step should work in such a way as to minimize the need for processing at the concatenation step.

Though we use 12 principal components for synthesis, problems with the first principal component are most important, since this component accounts for more than 57% of the variance of the final animation; furthermore, a discontinuity problem in the first component is a good predictor of discontinuity problems in the subsequent components. Given the results of selection, we scan across all the selected diphones and count the number of times the gap between boundaries of adjacent diphones, with regard to the first component, exceeds half the standard deviation of the component, as calculated throughout our whole database. We emphasize the fact that this operation takes place *before* concatenation. For later reference, let us call this number *visual metric*.

In exactly the same manner, we introduce metrics regarding the continuity of the first derivative of the visual trajectory (dP_i/dt) and of the fundamental frequency ($F0$). For the visual derivative, we count the number of times the gap between boundaries of adjacent diphones exceeds half the standard deviation of the derivative (*derivative visual metric*). For fundamental frequency, we count the number of times the gap between boundaries of adjacent diphones exceeds 0.25 Barks (*fundamental frequency metric*).

Finally, we define a fourth metric to assess the correctness of the rhythm structure of the synthesized utterance. For each vowel in the utterance, we measure the ratio of its duration to the sum of the durations of all vowels in the utterance. Then, we measure the same ratios in recorded equivalents of the same utterances. Thus, we define the *rhythm structure metric* as the number of vowels for which the value of the ratio calculated for the synthesized utterance is more than 150% of the value of the ratio calculated for the recorded utterance. This measure is quite different from the previous ones because it depends on a pre-recorded test corpus. The goal of this measure is not to obtain at the end the same rhythm structure but in fact to avoid big inconsistencies in the structure of the synthesized sentence.

4. Optimization of Weights

In order to fine-tune the weights involved in Eq. (1) we used a set of 20 short test sentences, recorded alongside the main corpus already presented.

We constructed a grid of values for the weights involved in Eq. (1). Since what is important for the Viterbi algorithm is not the absolute values of the weights but the ratios between them, we considered the weight w_{tc} as fixed to the value of 1. The weights w_{jc} and w_{vc} were given values at the range between 0 and 1 with a step of 0.1. The weight w_{dvc} was given values at the range between 0 and 0.2 with a step of 0.02. The choice of this difference in ranges and steps reflects the fact that continuity of the derivative of the visual trajectory is, in principle, less important to the final result compared to the continuity of the visual trajectory itself.

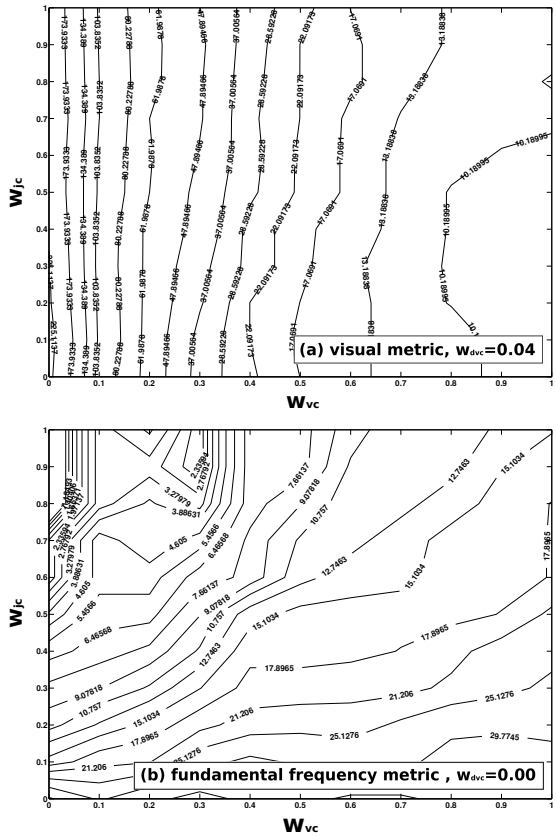

Figure 2: *(a) 2D contour graph of values of visual metric measured on the 20 test sentences, as a function of w_{vc} and w_{jc} when w_{dvc} is fixed to the value of 0.04. (b) 2D contour graph of values of fundamental frequency metric measured on the 20 test sentences, as a function of w_{vc} and w_{jc} when w_{dvc} is fixed to 0.*

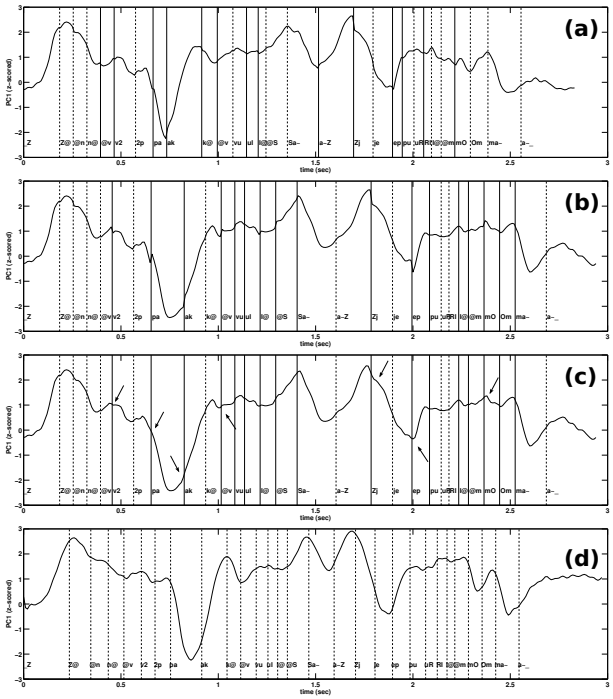

Figure 3: *First visual component (in z-scored units) for the test sentence "Je ne veux pas que vous le changiez pour le moment." (a) Synthesized using non-optimized weights, without processing at the visual joins. (b) Synthesized using optimized weights, without processing at the visual joins. (c) Synthesized using the optimized weights, after processing visual joins. Note the corrected details marked with arrows. (d) Original recorded trajectory. Horizontal axes denote time in seconds. 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.*

We synthesized the 20 test sentences and calculated the sums of the four metrics described in the previous section over these sentences, for every combination of weights in the grid. We plotted several contour graphs like the ones shown in Fig. 2. Since it was not imaginable that the observed minimum values of the four metrics could be achieved concurrently using a single combination of weights, we identified a set of target values that possibly could. These target values were slightly above the minimum values. We then normalized the four metrics by the target values and summed them up into a single merged metric. This merged metric quantifies how much the metrics deviate from the desired target values.

We used the Nelder-Mead algorithm [16], which is a well-defined commonly used nonlinear optimization technique based on the concept of a simplex, to minimize this merged metric over the four weights. In order to this, we developed a computational setup where the 20 test sentences were synthesized in every iteration of the algorithm. The algorithm gave us a minimum at the point $\{w_{tc}, w_{jc}, w_{vc}, w_{dvc}\} = \{1, 0.943, 0.897, 0.046\}$. This was a considerably different set than the set $\{1, 0.33, 3.33, 0\}$ which we used previously [7].

Fig. 3 shows synthesis results using these two sets of weights, for a single utterance from the test set. Though the trajectories corresponding to the (a) old and (b) optimized sets of weights look similar, it is important to notice that several of the diphones selected are different, a fact that is also reflected on the corresponding acoustics. The application of visual join processing in (c) further removes some small problems with the visual trajectories and produces a smoother animation. The synthesized trajectories bear a significant resemblance to the recorded trajectory for the same utterance, presented in (d). To illustrate the final result of our talking head synthesis system, Fig. 4 shows a series of faces corresponding to the trajectory in Fig. 3(c). Preliminary informal listening and seeing tasks gave the impression that the synthesized animations using the optimized weights were slightly improved. Nevertheless, perceptual tests need to be designed to provide an objective assessment.

5. Concluding Remarks

In acoustic-only speech synthesis, the problem of automatic fine-tuning of weights typically considers the relative weighting between two costs: target cost and join cost. An important step toward this goal is to link objective measurements with the subjective perception of the synthesis result by human listeners. It is a problem yet unsolved in a fully satisfactory way. Perhaps as an indication of disappointment, it has been argued that such attempts are unnecessary and there is no need to seek al-

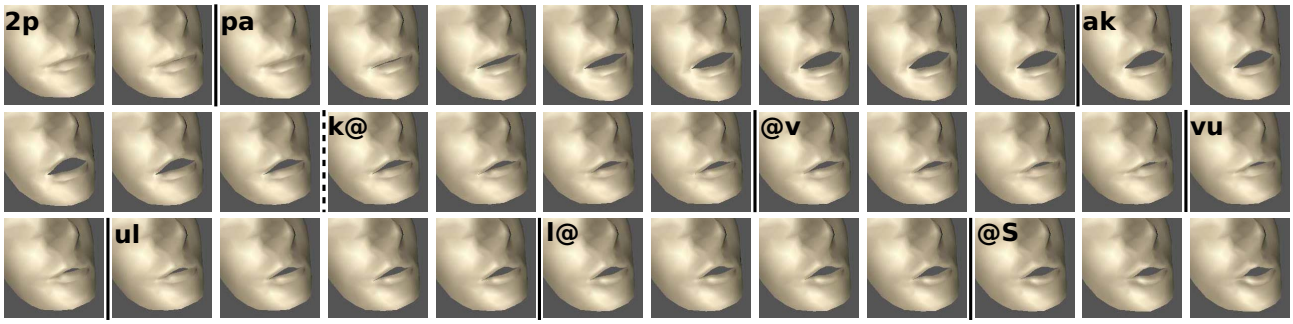

Figure 4: *Sequence of images, derived from synthesized 3D facial information, corresponding to part of the trajectory in Fig. 3(c), i.e. synthesized using the optimized weights, after processing visual joins. The words depicted are "pas que vous le". For sake of clarity, one image for every 20 ms is shown. Bars mark the boundaries between concatenated diphones (dashed when diphones are consecutive in the corpus).*

ternatives to manual fine-tuning of weights [17]. After all, the complexity is such that manual fine-tuning is not prohibitively expensive: all that needs to be determined is the ratio between the target cost weight and the join cost weight.

But in a bimodal unit-selection audiovisual synthesis setup, as in our case, the problem is much more complex. As seen in Eq. (1) two more join costs, to ensure continuity and smoothness in the visual domain, are added to the typical (acoustic) join cost. Moreover, these new visual costs are each constructed using 12 weights, introduced in Eq. (2), applied to the 12 principal components of visual information. For the selection of the latter set of weights, we resorted to a reasonable solution provided in the literature (weighting each component by its eigenvalue). In the work presented in this paper we targeted only the weights of Eq. (1).

Though the four metrics we introduced reflect perceptual considerations, we need more perceptual experiments to assess their relative importance, and thus to determine the exact way we should combine them into a single merged metric. Nevertheless, the experimental setup that we have developed can easily allow the incorporation of different merged metrics.

Will the final synthesized face animations be considerably improved after such an effort? Our first results indicate so, and we believe it is a question worth further investigation.

6. Acknowledgments

Our work was supported by the French National Research Agency (ANR - ViSAC - Project N. ANR-08-JCJC-0080-01). Our implementation of the Nelder-Mead algorithm was based on the Java library provided by Dr Michael Thomas Flanagan at www.ee.ucl.ac.uk/∼mflanaga. The 3D data used in this project were acquired by Brigitte Wrobel-Dautcourt and Marie-Odile Berger (Magrit group).

7. References

[1] W. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, p. 212, 1954.

[2] I. Pandžić, J. Ostermann, and D. Millen, "User evaluation: synthetic talking faces for interactive services," *The Visual Computer Journal*, vol. 15, pp. 330–340.

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

[4] G. Bailly, O. Govokhina, F. Elisei, and G. Breton, "Lip-Synching Using Speaker-Specific Articulation, Shape and Appearance Models," *EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[5] L. Wang, X. Qian, H. W., and S. F., "Synthesizing Photo-Real Talking Head via Trajectory-Guided Sample Selection," in *Interspeech*, Makuhari, Japan, 2010.

[6] 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.

[7] 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," in *Interspeech*, Makuhari, Japan, 2010.

[8] P. Taylor, *Text-to-speech synthesis*. Cambridge Univ. Press, 2009.

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

[10] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP*, Atlanta, USA, 1996.

[11] V. Strom and S. King, "Investigating Festival's target cost function using perceptual experiments," in *Interspeech*, Brisbane, Australia, 2008.

[12] F. Alías, L. Formiga, and X. Llorá, "Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms: A proof-of-concept," *Speech Communication*, vol. 53, no. 5, pp. 786 – 800, 2011.

[13] 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.

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

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

[16] J. Lagarias, J. Reeds, M. Wright, and P. Wright, "Convergence properties of the Nelder-Mead simplex method in low dimensions," *SIAM Journal on Optimization*, vol. 9, no. 1, pp. 112–147, 1999.

[17] R. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317 – 330, 2007.