Comparison between two predicting methods of labial coarticulation

Vincent Robert, Jacques Feldmar, Yves Laprie

Comparison between two predicting methods of labial coarticulation

Vincent Robert, Jacques Feldmar, Yves Laprie

Speech Group, LORIA UMR 7503
BP 239 - 54506 Vandoeuvre Les Nancy
FRANCE
http://parole.loria.fr
E-mail: vrobert@loria.fr

Abstract

The construction of a highly intelligible talking head involving relevant lip gestures is especially important for hearing impaired people. This requires realistic rendering of lip and jaw movements and thus relevant modeling of lip coarticulation. This paper presents the comparison between the Cohen & Massaro prediction algorithm and our concatenation plus completion strategy guided by phonetic knowledge. Although results show that Cohen & Massaro perform slightly better, the concatenation and completion strategy approximates consonant clusters markedly better particularly for the protrusion parameter. These results also show the concatenation and completion strategy could be easily improved via the recording of better reference models for isolated vowels.

1 Introduction

Coarticulation gives rise to articulatory reorganizations and gestures to enable the acoustic realization of a discrete sequence of phonemes with physical articulators. Unlike retentive coarticulation for which there is a broad consensus about the origin, i.e. mainly inertia of articulators, anticipatory and magnitude of anticipatory coarticulation are still debatable. The Look-ahead model proposed by Henke [7] addresses the issues from a purely phonological point of view. In accordance to his model, in a sequence V_1CCV_2 where V_1 is a non rounded vowel and V_2 is a rounded vowel, the protrusion movement begins as soon as possible, i.e. as early as the end of V_1, if intermediate consonants do not prevent this movement. On the contrary, the time locked model proposed by Bell-Berti and Harris [2] rests on the principle of overlapping gestures. Between these two models the hybrid model of Perkell and Chiang [11] decomposes anticipation into two phases: the first mainly depends on phonological constraints and the second one, faster, depends on the dynamics of articulators involved in coarticulation. Later, Abry and Lalouache [1] developed the expansionist theory which predicts that the anticipation movement can expand more easily than it can compress, and that anticipation is highly speaker dependent.

In parallel to these studies about anticipation, other computational models have been developed to describe the whole coarticulation phenomenon and not only anticipation. That developed by Öhman [8] targets tongue movements. It suggests that fast consonants gestures are superimposed on slow movements of the tongue dorsum required to realize vowels. The model of Cohen & Massaro [5] based on dominance functions proposed by Löfqvist [9] turns out to generate coarticulation patterns not very far from those generated by the time-locked model. Cohen and Massaro utilize exponential functions to represent the dominance functions attached to each phoneme and articulator. The sum of these functions produces the profiles of articulatory movements. Similarly to other domains of automatic speech processing neural maps and hidden Markov models [14, 15] have also been used with the difficulty of requiring sufficiently large training corpora. Since such corpora are not easily available, rule based systems, that of proposed by Pelachaud [10] and Beskow [3] for instance, exploit phonetic or articulatory rules to predict anticipation and/or retention articulatory gestures. Their behavior is not very far from the Look-Ahead approach.

Beskow [4] compared merit of several experimental models and noticed that the model of Cohen & Massaro gave the lowest Root Mean Squared Error

(RMSE) rate together with the highest correlation rate. Recently we proposed a new model to predict coarticulation which is capable of capturing relevant coarticulation information from a training corpus so as to reconstruct articulatory lip and jaw movements for any sequences of phonemes via a concatenation and completion strategy. Our approach described in [12, 13] combines two stages:

- one algorithm to predict the existence of coarticulation, which exploits standard articulatory-phonetic representation of speech sounds. This symbolic algorithm pilots the training of a numerical modeling of coarticulation patterns via sigmoid functions for each of the four articulatory parameters (lip protrusion, opening and stretching, jaw opening).
- a concatenation plus completion algorithm that synthesizes the dynamics of the four articulatory parameters from sigmoids trained on a corpus of audiovisual speech.

We present here the performances of this method compared to that of Cohen & Massaro.

2 Acquisition of articulatory data and synthesis strategy

A female speaker used to speak for deaf children recorded a large corpus which has been exploited to train our coarticulation model. This corpus is made up of all the vowels, semi-vowels and CV, all VCV combinations with V in /a,i,u/ , the most frequent 70 french VCCV and 100 phonetically balanced sentences. The recording consisted in acquiring the 3D positions of 190 markers (with a high density in the region of lips) painted onto the speaker's face and the speech signal. The stereo acquisition has been realized through two CCD cameras at the rate of 120 images per second.

30 sentences out of the 100 have been removed from this corpus to assess the synthesis of coarticulation. For each of the 30 sentences, the time evolution of lip protrusion, opening and stretching, as well as jaw opening, are synthesized either via the Cohen & Massaro approach [5], or our concatenation and completion algorithm [12, 13].

3 Global analysis of results

In order to enable the comparison between the concatenative and the Cohen and Massaro strategies, we implemented the Cohen and Massaro algorithm.

The measures of RMSE (Root Mean Squared Error) expressed in percentage with respect to the global range of the parameter considered, and correlation between real and synthetic data are given in table 1. We present two evaluations, a general one and a second one (see section 4) focused on difficult sequences which initially justified the design of our concatenative and completion algorithm.

It turns out that the implementation of the Cohen & Massaro algorithm gives slightly better results than the concatenation method. It should be noted that Beskow, with a similar corpus for Swedish, obtained results not as good as ours. This probably originates in the slight improvements implemented in our version of this algorithm.

However, the difference is less marked for protrusion. We also noticed that the amplitude registration intended to ensure the syntagmatic consistency of the labial parameters is more efficient for protrusion which presents a larger degree of freedom. On the other hand, the paradigmatic axis seems to be more important for other labial parameters and the Cohen & Massaro algorithm implicitly favors the paradigmatic axis since it is based on dominance functions attached to each of the phonemes. This probably explains that performances are better for lip opening and stretching, and jaw opening.

Table 1: *Comparison between the Cohen & Massaro prediction algorithm and our concatenation method.*

		Pro	Open	jaw	stre
Corr	C & M	77.61	85.86	85.33	85.58
	Conc	73.62	74.44	73.99	75.57
RMSE	C & M	8.37	6.48	7.62	6.79
	Conc	10.32	9.3	9.91	9.28

Generally speaking, even if statistical results are good, movements generated by the method of Cohen & Massaro presents some weaknesses. On the one hand some movements are more abrupt than those obtained via concatenation and real articulators would not be able to realize these movements. Figure 1 shows the synthesis results for protrusion on the sentence "Une galette pour jeudi" for both methods. Even if statistical results are very close for this sentence, several abrupt movements would probably give rise to some negative perceptive effects.

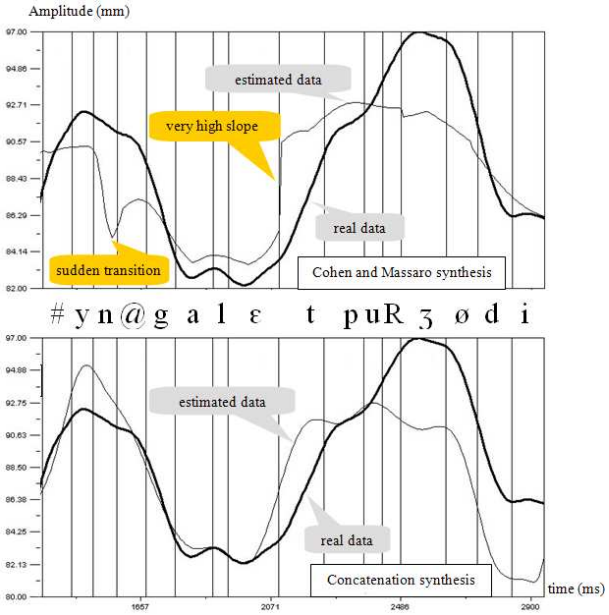

Figure 1: *Synthesis with the two methods of the sentence "Une galette pour jeudi"*

4 A finer analysis of results

A more intensive analysis shows that the Cohen & Massaro algorithm is not systematically better than ours; indeed some sentences are better synthesized with our approach. On the other hand it is more sensitive to the poor acquisition of some isolated phonemes. Indeed, isolated vowels are used to interpolate coarticulation patterns of VCVs not present in our corpus. The success of the interpolation thus depends on the quality of the recording of these isolated vowels. We noticed that some vowels, particularly /ɛ̃/, were not recorded correctly, what in consequence leads to poor error and correlation rates. The Cohen & Massaro approach which performs a global optimization of coarticulation parameters prevents this kind of weakness. Several occurrences of isolated vowels will thus be recorded in the future.

Our method gives results equivalent with those of Cohen & Massaro for protrusion. They are even better for phonemes characterized by a very small or a strong protrusion (especially /i,a,u,ʃ/) as shown by Figure 2. With respect to the realization of the bilabial closure it turns out that our implementation of the algorithm of Cohen & Massaro gets good results without further modification of the algorithm. This somewhat contradicts observations of Cosi [6].

Most of our corpus comprises VCV sequences. However, many consonant clusters exist in French

		Pro	Open	jaw	stre
Corr	C & M	57.17	87.19	88.68	89.56
	Conc	78.40	88.63	88.74	91.04
RMSE	C & M	14.81	6.70	7.50	7.51
	Conc	10.59	6.79	8.06	6.60

Table 2: *Comparison between the two predicting methods for the VCCV of our large corpus*

and we thus added the most frequent 70 VCCV in our corpus. In order to carry out the evaluation of coarticulation for consonant clusters we removed them from the training corpus. Table 2 exhibits results obtained by the two algorithms. They are very close for lip opening and stretching, and jaw opening as well. On the other hand, our algorithm gets markedly better results for protrusion (20% more for the correlation and 4% less for the RMSE).

These results show that dominance functions used by the algorithm of Cohen & Massaro are appropriate to approximate coarticulation when its scope is limited to one sound. On the other hand, coarticulation phenomena spreading over a longer phonetic context, i.e. especially those observed for more complex sequences as VCCV, are better approximated by our strategy that uses symbolic phonetic knowledge to predict the rough coarticulation pattern.

Our experiments also show that syntagmatic consistency is probably more important for the protrusion parameter than for other labial parameters. Indeed, the registration of protrusion amplitude applied to synthesized utterances to ensure the overall consistency has a positive impact only for protrusion.

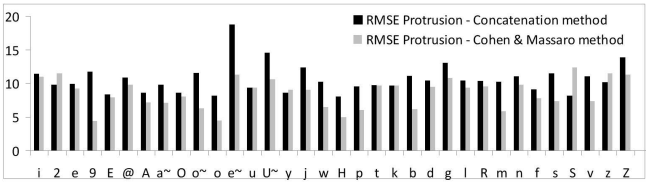

Figure 2: *RMSE for protrusion (SAMPA phonetic alphabet)*

5 Conclusion

This assessment shows that our implementation of the Cohen & Massaro algorithm performs slightly better than the concatenation of sigmoid functions. However, to a large extent this difference is explained

by reference targets of isolated vowels used by the completion algorithm to derive coarticulation patterns of VCV not present in the corpus. Recording several occurrences of the isolated vowels would enable a more consistent set of references to be built, and consequently better results to be obtained.

Another improvement concerns the symbolic prediction algorithm. The advantage, exhibited by results obtained for consonant clusters, is to get a first guess about the shape of coarticulation. However, this phonetic knowledge could probably be slightly improved and complemented. Indeed, the expected neutrality of some consonants with respect to labial coarticulation should be revisited. It should be noted that our approach enables this iterative refinement easily since only some sigmoid functions have to be added.

Finally, this evaluation focuses on numerical figures of merit not necessarily rendering the perceptive merit of coarticulation algorithms. We are now preparing a perception experiment to measure the real perceptive impact of these two approaches of coarticulation.

References

[1] C. Abry and T. Lallouache. Le MEM : un modèle d'anticipation paramétrable par locuteur: Données sur l'arrondissement en français. *Bulletin de la communication parlée*, 3(4):85–89, 1995.

[2] F. Bell-Berti and K. S. Harris. A temporal model of speech production. *Phonetica*, 38:9–20, 1981.

[3] J. Beskow. Rule-based visual speech synthesis. In *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech'95)*, pages 299–302, Madrid, Spain, 1995.

[4] J. Beskow. Trainable articulatory control models for visual speech synthesis. *International Journal Of Speech Technology 7*, pages 335–349, 2004.

[5] M. Cohen and D. Massaro. Modeling coarticulation in synthetic visual speech, 1993.

[6] P. Cosi and al. Labial coarticulation modeling for realistic facial animation. In *Proceedings of ICMI'02, 4th International Conference on Mutlimodal Interfaces*, pages 505–510, Pittsburgh, PA, USA, 2002.

[7] W. L. Henke. *Dynamic articulatory model of speech production using computer simulation , Unpublished doctoral dissertation*. PhD thesis, MIT Cambridge, 1966.

[8] S. Öhman. Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 41:310–320, 1967.

[9] A. Löfqvist. *Speech as audible gestures*. Hardcastle, W.J. and Marchal, A. (eds). Dordrecht: Kluwer Academic Publishers, 1990.

[10] C. Pelachaud, N. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive science, 20(1)*, pages 1–46, 1996.

[11] J. Perkell and C. Chiang. Premininary support for a 'hybrid model' of anticipatory coarticulation. In *Proceedings of the XIIth International Congress of Acoustics*, Toronto: Canadian Acoustical Association, A3-6, 1986.

[12] V. Robert, Y. Laprie, and A. Bonneau. A phonetic concatenative approach of labial coarticulation. *INTERSPEECH*, 2007.

[13] 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. *AVSP*, 2005.

[14] K. Tokuda, T. Kobayashi, and S. Imai. Speech parameter generation from HMM using dynamic features. In *ICASSP 95*, pages 660–663, Detroit, MI, USA, 1995.

[15] E. Yamamoto, S. Nakamura, and K. Shikano. Speech to lip movement synthesis by HMM. In *AVSP 1997*, pages 137–140, 1997.