AUTOMATIC MODELING OF COARTICULATION IN TEXT-TO-VISUAL SPEECH SYNTHESIS

Bertrand Le Goff
Institut de la Communication Parlée
UPRESA 5009/ INPG/Université Stendhal
BP25X, 38000 GRENOBLE CEDEX
Tel. +33 (0)4 76 82 41 28, FAX: +33 (0)4 76 82 43 35, E-mail: legoff@icp.grenet.fr

ABSTRACT

We have developed a visual speech synthesizer from unlimited French text, and synchronized it to an audio text-to-speech synthesizer also developed at the ICP (Le Goff & Benoît, 1996). The front-end of our synthesizer is a 3-D model of the face whose speech gestures are controlled by eight parameters: Five for the lips, one for the chin, two for the tongue. In contrast to most of the existing systems which are based on a limited set of prestored facial images, we have adopted the parametric approach to coarticulation first proposed by Cohen and Massaro (1993). We have thus implemented a coarticulation model based on spline-like functions, defined by three coefficients, applied to each target in a library of 16 French visemes. However, unlike Cohen & Massaro (1993), we have adopted a data-driven approach to identify the many coefficients necessary to model coarticulation. To do so, we systematically analyzed an ad-hoc corpus uttered by a French male speaker. We have then run an intelligibility test to quantify the benefit of seeing the synthetic face (in addition to hearing the synthetic voice) under several conditions of background noise.

1. INTRODUCTION

There is valuable and effective information afforded by a view of the speaker's face in speech perception by humans. Visible speech is particularly effective when the auditory speech is degraded, because of noise, bandwidth filtering, or hearing-impairment, as shown for long in English (Sumby & Pollack, 1954[17]; Summerfield et al., 1989[18] ; Erber, 1969 [8]; Erber, 1975[9]), and more recently in French (Benoît et al., 1994[4]). Synthetic faces also increase the intelligibility of natural speech when the facial gestures and speech sounds are coherent (Le Goff et al., 1995[10]; Le Goff et al., 1996[11]) . Therefore, we may easily assume that synthetic faces enhance the intelligibility of synthetic speech. However, this goal can only be reached if the articulatory parameters of the facial animation signal the same message as the auditory speech (McGurk & MacDonald, 1976[15]).

Another main problem resides in coarticulation. The first visual synthesizers were based on visual speech units placed side by side, and thus did not reproduce the coarticulation which makes phonetic units overlap. In order to bridge the gap, two strategies have been adopted. The first one (*look-ahead model*, Kozhevnikov-Chistovich, 1965[13]) mainly reproduces anticipatory coarticulation for protrusion, and is based on C^nV cluster whose V unit sets its protrusion for the whole cluster. The second one (*time-locked model*, Bell-Berti & Harris, 1982[2]) makes V protrusion begin a certain amount of time prior the vowel. Bladon & Al-Bamerni[5] and then Perkell & Chiang [16] have used both methods to produce an *hybrid model.*

But the reason of using one of these three models comes from empirical results. Several studies (Lubker & Gay, 1982[14]; Boyce, 1990 [6]; Abry & Lallouache, 1991[1]) have shown that each model was language-dependent. For example, in French, Abry and Lallouache[1] lay stress on the fact that these models were not able to render early protrusion such as the one produced by /y/ in [istrstry].

This is the reason why Cohen and Massaro[7] used a more flexible general framework using dominance functions. We thus decided to use this spline-like approach. We extended this method and unlike them, we used a data-driven approach to identify the coefficients of our splines.

2. DESCRIPTION OF THE MODEL

2.1 Introduction

Splines are used to draw curves through several control points. Each point locally controls the curve through a dominance function which is used as a weighting function in sort of barycenter calculus of all control points.

The main problem with spline weighting functions is that each dominance function is built on the same model (same range of influence). Moreover, they are not very interpretable.

2.2 General formula of dominance function

Cohen and Massaro[7] used an exponential-based dominance function : $f(t) = \alpha \times e^{-\theta_i \cdot |t - t_0|}$. For a given parameter and a given viseme, α represented the dominance of this viseme over the others, θ_1 and θ_2 characterized the amplitude of respectively anticipatory

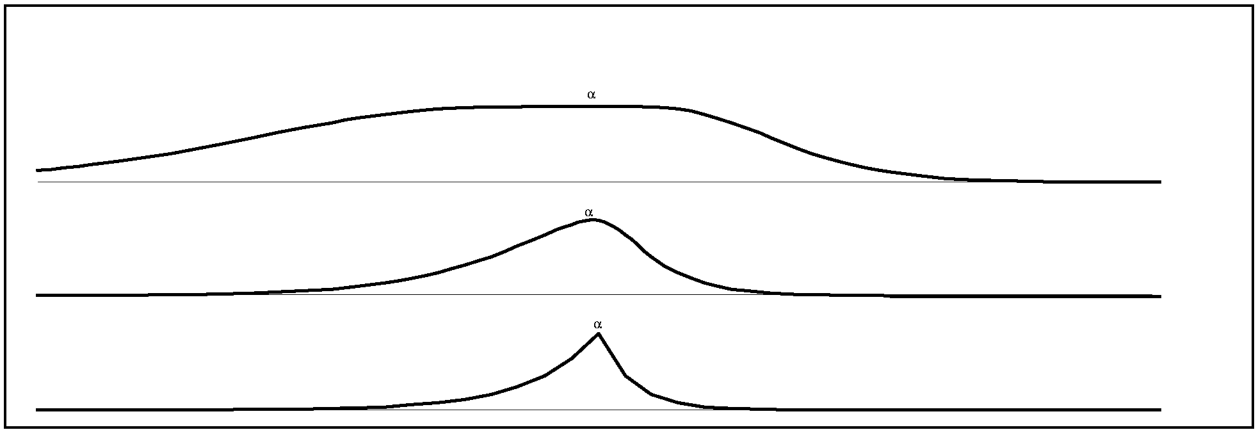

Figure 1 : asymmetrical dominance functions ($\theta_1 < \theta_2$) for several continuity orders. From bottom to left, C^0, C^1 and C^3 functions.

and carry-over coarticulation. We extended this formula so that the function is n-continue (C^n function) :

$$f(t) = \alpha \times e^{-\theta_i \cdot |t-t_0|} \times \sum_{j=0}^{n-1} \frac{\theta_i^j}{j!} \times |t-t_0|^j$$

with t_0 at the center of phoneme.

We obtained functions shown on Figure 1. They are used in a barycenter-like formula :

$$parameter_i(t) = \frac{\sum_k f_{k,i}(t) \times T_{k,i}}{\sum_k f_{k,i}(t)}$$

where $T_{k,i}$ is the target (or control point) whose $f_{k,i}$ is the dominance function. The barycenter formula took in account only a number of neighboring visemes which is known to be the furthest influence range; we have thus chosen a window of 14 visemes to respect the coarticulation for French (Abry & Lallouache[1]; Benguerel & Cowan[3]). But the corpus used to calculate data contained only 3 visemes in each sequence.

2.3 Control points, or visemes

We have defined 16 classes of visual phonemes (visemes) represented by eight parameters (five for the mouth : A width, B height, C lip contact protrusion, P1 upper lip protrusion, P2 lower lip protrusion; one for the chin : M vertical displacement; two for the tongue : Ta angle with the jaw, Tl length).

There are 7 classes of vowels : [i], [e, ɛ, ɛ̃], [a], [y, u, ø, o, õ], [ɔ], [œ, œ̃], [ã] and 9 classes of consonants : [b, p, m], [ʒ, ʃ], [t, d, n], [f, v], [s, z], [k, g], [l], [R], [w].

The seven vowel targets values T_{ij} of each parameter P_j (tongue parameters excluded) were directly measured on the quasi-stationary production of seven vocalic visemes V_i ={a, e, i, ã, O, œ, y} repeated several times in the sentence-like sequence "C'est pas $V_iV_iV_i$z ?".

Consonant targets were defined by an automatic analysis described below.

The 128 tongue coefficients and targets were estimated from introspection.

2.4 Automatic approach

For each visemic class and for each parameter, 4 coefficients were to be defined (α, θ_1, θ_2 and T) i.e. 470 variables (the 42 vowel targets -tongue excluded- were already defined).

The other 342 coefficients describing consonant targets and viseme coarticulation were estimated altogether from a thorough analysis of a corpus consisting of 9 repetitions of sentences of the form "c'est pas $V_1CV_2CV_1$z ?", where V_1 and V_2 were {a, i, y} and C was {b, d, g, ʒ, l, R, v, w, z}.

The extent of coarticulation was thus limited to V_1CV_2 and CVC triphone combinations. The six visible parameters were finally measured every 20 ms on each image field, and the centers of the acoustic realizations were hand-labeled.

Through an analysis-by-synthesis process, parameters simulated evolution was compared with human parameters evolution thanks to calculation of euclidian distances between trajectories. Then, relaxation method (optimization method using dichotomy applied successively on each coefficient of each visemic class) was used to minimize distances. Optimization was done independently on each parameter. For each parameter, several formulas (i.e. several continuity orders) were tested in order to find the one which best approximates natural trajectories. The order has been found to range from 1 to 3 in our analysis, depending on the parameter.

Finally, dominance characteristics of the other four visemic vocalic classes {e, ã, O, œ} were extrapolated from those observed with {a, i, y}.

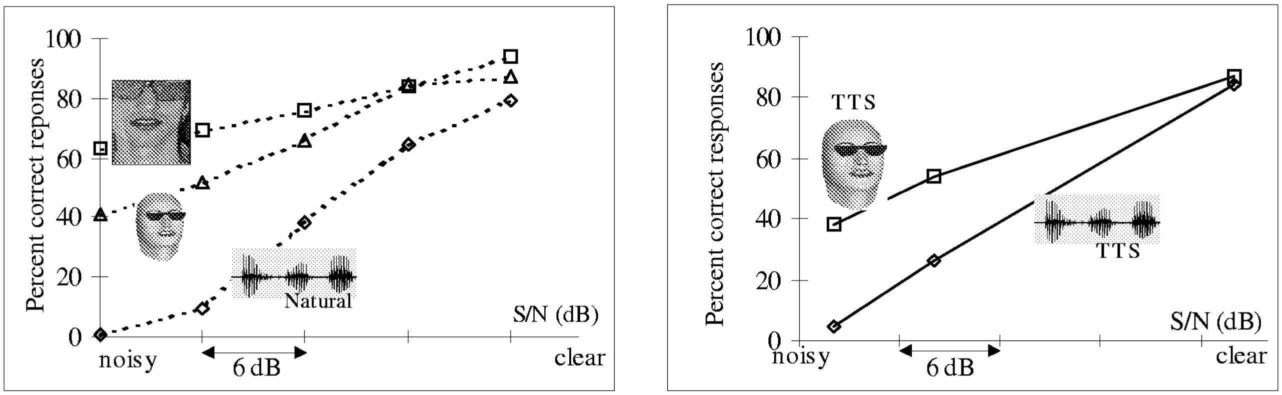

Figure 2 : Intelligibility scores in the identification of 18 stimuli, as a function of acoustic degradation, depending on the mode of presentation. Left, human face, face model animated by speaker and natural audio curves in dashed line from Le Goff et al. , 1995[10]. Right, synthetic audio and face animated by calculated parameters, in plain lines.

3. EVALUATION

We have evaluated the intelligibility of this new version of our synthesizer through an experimental protocol largely used in the past for audio-visual tests so that comparisons with previous data could be made. It involves the auditory only and audio-visual presentation of speech material to subjects under several conditions of acoustic degradation. This test has already been performed with entirely natural speech (Benoît et al., 1994[4]; Le Goff et al., 1996[11]) as well as with natural acoustic speech and a series of facial models, including that here used, animated from measurements of the speaker's facial gestures (Le Goff et al., 1995[10]; Le Goff et al., 1996[11]).

3.1 Protocol

Eighteen nonsense words were presented to French normal-hearers. The corpus was made of VCVCV sequences, with V=/a/, /i/ or /y/, and C=/b/, /ʒ/, /l/, /R/, /v/ or /z/.

Each word was generated under two conditions of presentation :

1) no video : the eighteen sentences, produced by the audio synthesizer, were degraded by addition of pink noise, at three S/N levels (-16 dB, -10 dB and +8 dB).
2) synthetic face : the face model (front view) was Gouraud shaded and animated at more than 25 ips on the 19 inch monitor of an SGI Elan. The same synthetic audio files as in 1) were synchronized with the face, under the three above conditions of noise.

Ten normal-hearing French subjects took part in the experiment. The 90 stimuli were presented in one sequence, with ten extra stimuli appended before the actual test so that subjects could adapt to the test conditions. Subjects answered through a mouse-driven interface. They were recommended to respond to both the vowel and the consonant, as much as they could guess it. A "?" response was tolerated, however.

3.2 Global intelligibility

A test word was first considered correct only if both the vowel and the consonant were correctly identified. Global results are reported in Figure 2 in plain lines, as well as results from previous studies in dashed lines (Le Goff et al., 1995[10]). Signal intensities are not fully comparable in part because of differences in sampling rate. Only relative comparisons are significant, this is the reason why graphics have not been superimposed.

Adding the video shows a dramatic gain in intelligibility : in noisy conditions, the synthesizer restores more than a third of the missing auditory information. In speech-reading condition (synthetic audio very noisy), it succeeds in restoring two thirds of the benefit that the natural face adds to the natural audio.

The face driven by synthetic parameters is thus almost as well recognized as the same face driven by parameters extracted from the speaker's face.

3.3 Consonant confusions

Consonant confusions are presented in Table 1 at S/N = – 10 dB where differences are at their maximum.

	b	ʒ	l	R	v	z	?
b	**50**	10	7	10	7	3	13
ʒ	10	20	20	10	13	7	20
l	17	7	**33**	17	13		13
R	25	3	3	**50**	3	3	13
v	23	10		10	23	13	21
z	**23**	17	3	3	20	20	14

	b	ʒ	l	R	v	z	?
b	**87**			3	7	3	
ʒ	3	**58**	13	13	10	3	
l			**94**	3		3	
R	13	7	17	**47**	3	13	
v	**40**	10	13	7	20	10	
z	7	27	10	10	3	**43**	

Table 1 : Confusion matrices of consonants, irrespective of the response on the vowel (S/N = – 10 dB). Stimuli are presented in rows. Percepts are presented in columns. Scores are in percent. Top, TTS audio; bottom, AV synthesizer.

Except for /v/, there is a very strong disambiguation due to visual information :

- /b/ is well recognized, although it is given as a response to many /v/ stimuli.
- /ʒ/ identification is improved thanks to the strong protrusion of the lip contact.
- /l/ is the consonant best identified audio-visually thanks to the tongue movement.
- /R/ identification is not improved when vision of the face is added to audio. This is obviously due to its lack of visual characteristic.
- /z/ is disambiguated by the vision of the face, whereas it is not recognized at all auditorily.

3.4 Vowel confusions

Vowel confusions are presented in Table 2 at S/N = – 10 dB where differences are at their maximum.

	a	i	u	?
a	**98**			2
i	7	**43**	38	12
u	2	28	**67**	3

	a	i	u	?
a	**100**			
i	15	**68**	17	
u			**100**	

Table 2 : Confusion matrices of consonants, irrespective of the response on the consonant (S/N= – 10 dB). Stimuli are presented in rows. Percepts are presented in columns. Scores are in percent. Left, TTS audio; right, AV synthesizer.

Vision of the face allows a total disambiguation of all vowels, even if some of /i/ stimuli remains slightly mixed up with /a/ and /y/.

4. CONCLUSION

We have presented a new version of our audiovisual speech synthesizer from unlimited French text. The protocol we have here exposed can be used for modeling any language. The whole system (audiovisual synthesizer + face) can be used as a human/machine interface, and would also be useful in computer-aided learning of foreign-language or of speech-reading. Moreover, it allows the generation of audio-visual stimuli totally controlled, and can be used to investigate bimodal speech perception.

5. REFERENCES

[1] C. Abry and M.T. Lallouache, "*Audibility and stability of articulatory movements: Deciphering two experiments on anticipatory rounding in French*", Proceedings of the XIIth International Congress of Phonetic Sciences, Vol. 1, pp. 220-225, Aix-en-Provence, France, 1991.

[2] F. Bell-Berti, and K.S Harris, "Anticipatory coarticulation : Some implications from a study of lip rounding ", *Journal of the Acoustical Society of America*, Vol. 65, pp. 1268-1270, 1982.

[3] A.P. Benguerel and H.A. Cowan, "Coarticulation of upper lip protrusion in French", *Phonetica*, Vol. 30, pp. 41-55, 1974.

[4] C. Benoît, T. Mohamadi and S. Kandel, "Audio-visual intelligibility of French speech in noise", *Journal of Speech & Hearing Research*, Vol. 37, pp. 1195-1203, 1994.

[5] R.A. Bladon and A. Al-Bamerni, "One stage and two-stage temporal patterns of velar coarticulation", *Journal of the Acoustic Society of America*, Vol. 72, S104(A), 1982.

[6] S.E. Boyce , "Coarticulation organization for lip rounding in Turkish and English", *Journal of the Acoustical Society of America*, Vol. 88, pp. 600-607, 1990.

[7] M.M. Cohen and D.W. Massaro, "Modeling Coarticulation in Synthetic Visual Speech", *Models and Techniques in Computer Animation*, N. M. Thalmann & D. Thalmann (Eds.), Tokyo: Springer-Verlag, pp. 139-156, 1993.

[8] N.P. Erber, "Interaction of audition and vision in the recognition of oral speech stimuli", *Journal of Speech & Hearing Research*, Vol. 12, pp. 423-425, 1969.

[9] N.P. Erber, "Auditory-visual perception of speech", *Journal of Speech & Hearing Research*, Vol. 40, pp. 481-492, 1975.

[10] B. Le Goff, T. Guiard-Marigny and C. Benoît, "*Read my lips ... and my jaw ! How intelligible are the components of a speaker's face ?*", Proceedings of Eurospeech'95, Madrid, Spain, Vol. 1, pp. 291-294, 1995.

[11] B. Le Goff, T. Guiard-Marigny and C. Benoît, "Analysis-Synthesis and Intelligibility of a Talking Face", *Progress in speech synthesis*, J.P.H. Van Santen, R.W. Sproat, J.P. Olive & J. Hirschberg Editors, Springer Verlag New York, pp. 235-246, 1996.

[12] B. Le Goff and C. Benoît, "*A text-to-audiovisual-speech synthesizer for French*", Proceedings of the 4th International Conference on Spoken Language Processing, Philadelphia, PA, USA, Vol. 4, pp. 2163-2166, 1996.

[13] V.A Kozhevnikov and L.A. Chistovich, "*Rech:artikulyatsiya i Vospriyatiye (Moscow-Leningrad, 1965). Trans. Articulation and perception*", Washington, D.C. : Joint Publication Research Service, Vol. 30, pp. 543, 1965.

[14] J. Lubker and T. Gay, "Anticipatory labial coarticulation: experimental, biological, and linguistic variables", *Journal of the Acoustical Society of America*, Vol. 71, pp. 437-448, 1982.

[15] H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices", *Nature*, Vol. 264, pp. 746-748, 1976.

[16] J.S. Perkel. and C. Chiang,, "*Preliminary support for a "hybrid model" of anticipatory coarticulation*", Proceedings of the 12th International Conference of Acoustics, A3-6, 1986.

[17] W.H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise", *Journal of the Acoustical Society of America*, Vol. 26, pp. 212-215, 1954.

[18] Q. Summerfield, A. MacLeod, M. McGrath and M. Brooke, "Lips, teeth, and the benefits of lipreading", *Handbook of Research on Face Processing*, A.W. Young and H.D. Ellis Editors, Elsevier Science Publishers, pp. 223-233, 1989.