A TEXT-TO-AUDIOVISUAL-SPEECH SYNTHESIZER FOR FRENCH

Bertrand Le Goff, Christian Benoît

Institut de la Communication Parlée, INPG/ENSERG-Université Stendhal, BP25X
38040, GRENOBLE Cédex 9, France

ABSTRACT

An audiovisual speech synthesizer from unlimited French text is here presented. It uses a 3-D parametric model of the face. The facial model is controlled by eight parameters. Target values have been assigned to the parameters, for each French viseme, based upon measurements made on a human speaker. Parameter trajectories are modeled by means of dominance functions associated with each parameter and each viseme. A dominance function is characterized by three coefficients so that coarticulation finally depends on the phonetic context, the speech rate, and an "hypo-hyper articulation" coefficient adjustable by the user. Finally, the visual and audiovisual intelligibility of our visual synthesizer has been evaluated in its first version, and compared to that of the acoustic synthesizer on which it was implemented.

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[18]; Summerfield et al., 1989[19]; Erber, 1969[7]; Erber, 1975[8]), and more recently in French (Benoît et al., 1994[3]). Synthetic faces also increase the intelligibility of natural speech when the facial gestures and speech sounds are coherent (Le Goff et al., 1994[12]; Le Goff et al., 1995[13]). 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. If an ambiguous message or even a contradictory message is given by the visible speech, then intelligibility is decreased (McGurk & MacDonald, 1976[14]).

Most of the existing parametric models of the human face have been developed to optimize the visual rendering of facial expressions (Parke, 1974[15]; Waters, 1987[20]; Platt & Badler, 1981[16]; Viaud & Yahia, 1992[21]). As regards the visual synthesis of speech, existing systems are essentially based on a limited set of facial images occurring in the natural production of speech that are displayed one after the other, depending on the phoneme simultaneously uttered by the acoustic synthesizer (e.g., Saintourens et al., 1990[17]; Henton et al., 1994[11]). Actually, coarticulation effects and transition smoothing are much more naturally simulated by means of parametric models specially controlled for speech production, as that developed by Cohen and Massaro (1993)[6] from the original Parke (1974)[15]'s model. In that perspective, a high-resolution model of the lips, controlled by only five parameters, was developed at the ICP (Guiard-Marigny et al., 1994[10]) and it was then implemented by Le Goff et al.(1994)[12] onto Cohen & Massaro (1990)[5]'s face model which we modified so that it is now controlled by only eight articulatory parameters. Following the strategy adopted by Cohen and Massaro (1993)[6], we have assigned target values and dominance functions to each of the eight parameters caracteristic of each French viseme. This technique allows parameter trajectories to be smoothed over time. It thus gives a fair account of the coarticulation phenomenon. Our visual synthesizer is now synchronized with the diphone-based speech synthesizer developed at the ICP (Bailly & Guerti, 1991[1]). Finally, a first evaluation of the system performance was run in terms of its visual and audiovisual intelligibility compared to the auditory intelligibility of the acoustic synthesizer.

2. TALKING FACE SYNTHESIS

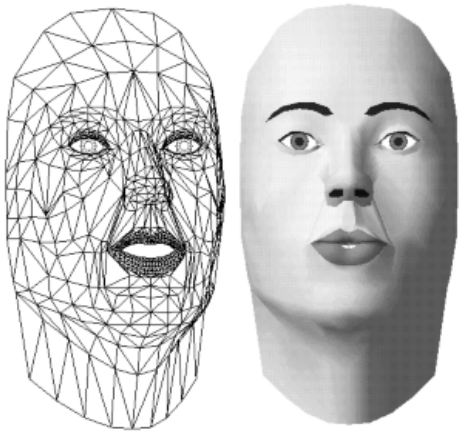

Figure 1: The ICP facial model. Left: The wire-framed underlying structure (without teeth and tongue) uttering a /i/. Right: A Gouraud-shaded view of the model uttering a /y/.

The talking face synthesizer uses a 3-D face model controlled by eight parameters (see Figure 1). From an extended corpus, Benoît et al. (1992)[2] identified twenty or so mouth shapes that best describe the structure of lip and jaw gestures in French. We used the same corpus to select a set of eight parameter targets characteristic of each French viseme. Trajectories in-between these targets are calculated through mathematical rules based on dominance functions. The dominance functions are defined by three coefficients (absolute amplitude, left and right influence across time) for one given parameter and one given viseme. All coefficients were calculated from an extensive analysis of the above mentioned reference corpus. Our facial model is ultimately

animated from a sequence of visemes and their associated duration. These data are easily obtained from a string of phonemes and their duration, as predicted and provided by a text-to-speech synthesizer.

2.1. The Face Model

Our facial model (see figure 1) was built up upon the 3-D lip model developed at the ICP (Guiard-Marigny et al., 1994[10]). This 3-D model was implemented by Le Goff et al. (1994)[12] on a version of the original Parke (1974)[15]'s model later modified by Cohen & Massaro (1990)[5]. The complete model of the face is controlled by eight parameters: five for the lips (horizontal inner width, vertical inner height, lip contact protrusion, lower and upper lip protrusion), one for the chin and two for the tongue (pitch angle and horizontal lengthening).

2.2. Target Values

Seventeen "visemes" (Fisher, 1968[9]) are necessary to describe the visual production of French at the symbolic level: [a], [i], [y, u, ø, o, õ], [e, ɛ, ɛ̃], [ɔ], [œ, œ̃], [ã], [p, b, m], [t, d, n], [k, g], [f, v], [s, z], [ʃ, ʒ], [l], [ʁ], [w], [j]. Two classes were added: a "prephonatory" shape, with lips ajar; and a "rest" position, with closed lips, because these shapes do not match "speech" gestures. For each of the 19 classes, target values were assigned to the eight parameters from measurements on a reference French speaker, filmed from front and profile, uttering the vowels in a steady state and the consonants in a /a/, /i/, and /y/ context. For consonants, the target values of the parameters were averaged on the three contexts.

2.3. Dominance Functions

The "dominance" of a viseme models the extent of its coarticulation effect on its neighbors. The dominance of a parameter is defined by three coefficients: α, θ_1, θ_2. For a given parameter and a given viseme, α reflects the dominance of this viseme on all others, θ_1 characterizes the amplitude of its anticipatory coarticulation, and θ_2 that of its carry-over coarticulation.

For each viseme, the 24 coarticulation coefficients were calculated from the analysis of parameter trajectories measured on our reference French speaker except for the tongue parameters which were simply predicted from phonetic knowledge.

The dominance function used is one of those proposed by Cohen and Massaro (1993)[6]. Its general form is:

$f = \alpha . e^{-\theta . |t_0 - t|} . (1 + \theta . |t_0 - t|)$ with t_0 corresponding to the acoustic center of the corresponding phoneme.

Except for inner lip height parameter which is adjusted by specific rules so lips remain closed during bilabial occlusion, the evolution of a parameter between two targets takes a sigmoïdal shape. Its final trajectory depends on the coefficients of the two targets (See Figure 2), and on time duration between them. The shorter the duration, the stronger the coarticulation, and the smoother the trajectory.

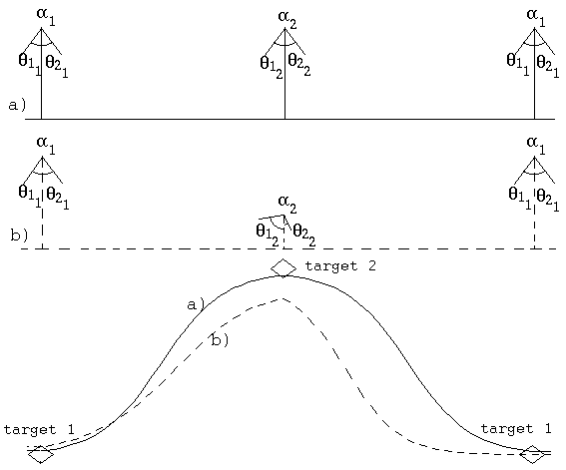

Figure 2: Schematic of parameter transitions from viseme 1 to viseme 1 through viseme 2. Upper pannel: The Dominance functions represented by their three coefficients. Lower pannel: The resulting parameter trajectory over theorical targets (diamonds).
The plain curve (a) was obtained from three targets with identical dominance. In this case, target 2 is reached.
The dotted curve (b) was obtained from targets with different dominance. An undershoot is then observed: Target 2 is not reached because α_2 is too small, i.e., this second viseme parameter has little coarticulatory effect on viseme 1. In addition, this parameter shows a stronger anticipatory effect from viseme 2 on the first viseme 1 (low θ_{1_2}) and a weaker carry-over effect on the second viseme 1 (high θ_{2_2}).

3. INTELLIGIBILITY OF THE SYSTEM

We have evaluated the intelligibility of this first 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[3]), as well as with natural acoustic speech and a series of facial models, including that here presented, animated from measurements of the speaker's facial gestures (Le Goff et al., 1995[13]).

3.1. Protocol

Eighteen nonsense words were presented to thirty 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 six conditions of presentation: A = audio alone at a fast rate; AV = audio-visual at the same fast rate; VF = visual alone at the same fast rate (5.5

syllables/second), VC = visual at a conversational rate (3.3 syllables/second). In addition, three other visual conditions have been tested and discussed elsewhere (Benoît et al., 1996[4]).

Subjects had to give a response to both the vowel and the consonant from a 18-choice response box on a screen. They also had to rank the confidence in their response on a 0-10 scale for each stimulus. Stimuli were generated in real-time at 25 frames per second on an SGI Indy XZ, so that the test was run individually and self-paced by the subject.

Screen was located 1 m from the subject. The actual size of the face was 14.5 cm high and 8.0 cm width on the screen.

30 subjects participated in the experiment. In the A and AV conditions, 19 subjects could hear the acoustic synthesis at a low level with S/N = 3 dB(A). The 11 other subjects could hear the acoustic synthesis at a comfortable level with S/N = 18 dB(A).

3.2. Results and discussion

First, a response was considered correct if both the vowel and the consonant were correctly identified. Global results are reported in table 1.

	A	AV	VF	VC
low	37	45	27	28
loud	68	77		

Table 1: mean proportion of correctly identified stimuli (in percent).

Whatever the loudness of the acoustic synthesizer, vision of the synthetic face significantly enhances the intelligibility of the synthesizer ($p<0.05$).

In the visual alone condition, scores are well above chance: 27.5% vs. 5.6%. Subjects could thus successfully speechread a fourth of the synthetic material visually presented.

Visual confusion of vowels is presented in Table 2. The French rounded /y/ is seldom confused with the other two vowels. However, /a/ and /i/ are mixed up in one third of the cases.

%	a	i	y
a	73	19	8
i	42	48	10
y	11	12	77

Table 2: Visual confusion of vowels, regardless of the response on the consonant. Stimuli are in rows, percepts are in columns. Scores are in percent

Visual confusion of consonants is presented in Table 3.

%	b	ʒ	l	r	v	z
b	60	7	12	8	10	3
ʒ	7	29	35	13	7	9
l	6	8	71	7	4	4
r	13	12	38	16	13	8
v	26	7	24	12	30	1
z	9	10	37	22	10	12

Table 3: Visual confusion of consonants, regardless of the response on the vowel. Stimuli are in rows, percepts in columns. Scores are in percent

- /b/ and /l/ are the consonants best identified visually. Lip closure is a robust cue to visual identification of bilabials. Tongue movement also helps subjects identify /l/, in all contexts. However, /l/ is frequently given as a response to /ʒ/, /R/ and /z/.
- /ʒ/ is mostly confused with the liquids.
- /R/ is largely mixed up with /l/ in /i/ and /y/ contexts.
- /v/ is somewhat confused with /b/, especially in /i/ context. This is probably due to the high dynamics of the lips in the two contexts. More surprisingly, /v/ is perceived as /l/ in 24% of the cases (/a/ and /y/ contexts).
- /z/ is very poorly identified. It is mostly confused with /l/ in the /i/ and /y/ contexts, and with /r/ in the /a/ context.

Overall speechreading scores were not significantly better at the conversational rate (28%), although duration was increased by 1.6. This last observation tends to prove that the coarticulation model used is rather robust to hypoarticulation, since identification of vowels and consonants is not significantly decreased when undershoots are generated, that is when targets are not reached.

However, synthesizer's scores are far below those previously obtained in [13]. The same corpus uttered by the same face (tongue excluded) animated by parameters measured on a human face was identified correctly at 41.3% from vision alone. As with a natural face, subjects could identify 63% of the stimuli.

Contrarily to /i/ and /y/, /a/ uttered by the synthesizer was better recognized than /a/ uttered by the synthetic face animated from natural gestures (72.8% vs. 58.3%).

All these results show that our model of coarticulation doesn't render adequately the naturalness of speech gestures, and progress should be made in the future, based on the results of this test, to improve the visual intelligibility of our system.

4. CONCLUSION

We have presented the first version of an audiovisual speech synthesizer from unlimited French text. It is a great tool to model and better understand how speech is produced by humans. It also allows the generation of highly controlled audio-visual stimuli that could not be produced by humans. Therefore, it may serve in many experiments to investigate visual and bimodal speech perception: It was recently used to evaluate the importance of *hypo-* and *hyper-speech* in the visual modality (Benoît et al., 1996[4]). In addition, audio-visual synthesis has many potential applications as a human/machine interface, and in computer-aided learning of a foreign languages and of speechreading.

The audio-visual speech synthesizer here presented is a first step towards a more efficient system. First, results from this intelligibility test will help us correct several details. Dominance functions should be refined through analysis of a larger corpus. Time position of parameter targets should be modified since they are located at the center of the acoustic realization of a phoneme in this first version. In addition, facial expressions, such as eye blinking, head nodding, eyebrow raising, etc., should be controlled from syntax and prosody.

Acknowledgments: this study was supported by the French CNRS and by the European ESPRIT-BRA project No 8579 "MIAMI". We are debtful to Michael Cohen and Dominic Massaro for having made available their parametric model of the face. We would like to thank Christian Abry, Ali Adjoudani, Angela Fuster-Duran, Thierry Guiard-Marigny and Jean-Luc Schwartz for technical or scientific support.

5. REFERENCES

1. Bailly G. & Guerti M. "Synthesis-by-rule for French", *Proceedings of the XIIth International Congress of Phonetic Sciences*, Aix-en-Provence, France, n°2, 506-511, 1991.
2. Benoît C., Lallouache M.T., Mohamadi T. & Abry C. "A set of French visemes for visual speech synthesis", *Talking Machines: Theories, Models and Designs*, G. Bailly and C. Benoît, Eds, Elsevier Science Publishers B.V., North-Holland, Amsterdam, 485-504, 1992.
3. Benoît C., Mohamadi T. & Kandel S. "Audio-Visual Intelligibility of French speech in noise", *Journal of Speech & Hearing Research*, n°37, 1195-1203, 1994.
4. Benoît C., Fuster-Duran A. & Le Goff B. "An investigation of hypo- and hyper-speech in the visual modality", *Proceedings of ETRW 96*, Autrans, France, 1996.
5. Cohen M.M. & Massaro D.W. "Synthesis of visible speech", *Behaviour Research Methods, Instruments \& Computers*, 22(2), 260-263, 1990.
6. Cohen M.M. & Massaro D.W. "Modeling coarticulation in synthetic visual speech", *Proceedings of Computer Animation93*, Magnenat-Thalmann & Thalmann Eds, Geneve, Suisse, 1993.
7. Erber N.P. "Interaction of audition and vision in the recognition of oral speech stimuli" *Journal of Speech & Hearing Research*, n°12, 423-425, 1969.
8. Erber N.P. "Auditory-visual perception of speech", *Journal of Speech & Hearing Disorders*, n°40, 481-492, 1975.
9. Fisher C.G. "Confusions among visually perceived consonants", *Journal of Speech & Hearing Research*, n°15, 474-482, 1968.
10. Guiard-Marigny T., Adjoudani A. & Benoît C. "A 3D model of the lips", *Proceedings of the 2nd ETRW on Speech Synthesis*, New Platz, USA, 1994.
11. Henton C. & Litwinowicz P. "Saying and seeing it with feeling: techniques for synthesizing visible, emotional speech", *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, New York, USA, 73-76, 1994.
12. Le Goff B., Guiard-Marigny T., Cohen M. & Benoît C. "Real-Time Analysis-Synthesis and Intelligibility of Talking Faces", *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, New York, USA, 53-56, 1994.
13. Le Goff B., Guiard-Marigny T. & Benoît C. "Read my lips ... and my jaw! How intelligible are the components of a speaker's face?" *Procedings of Eurospeech'95*, Madrid, Spain, 1995.
14. McGurk H. & MacDonald J. "Hearing Lips and Seeing Voices", Nature, 264, 746—748, 1976.
15. Parke F.I. "A parametric model for human faces", *PhD Dissertation*, University of Utah, Department of Computer Sciences, 1974.
16. Platt S.M. & Badler N.I. "Animating Facial Expressions", *Computer Graphics*, 15, 245-252, 1981.
17. Saintourens M., Tramus M.H., Huitric H. & Nahas M. "Creation of a synthetic face speaking in real time with a synthetic voice", *Proceedings of ESCA workshop on speech synthesis*, Autrans, France, 249-252, 1990.
18. Sumby W.H. & Pollack I. "Visual contribution to speech intelligibility in noise", *Journal of the Acoustical Society of America*, n°26, 212-215, 1954.
19. Summerfield Q., MacLeod A., McGrath M. & Brooke M. "Lips, teeth, and the benefits of lipreading", in *Handbook of Research on Face Processing*, A.W. Young & H.D. Ellis Editors, Elsevier Science Publishers, 223-233, 1989.
20. Waters K. "A Muscle Model for Animating Three-Dimensional Facial Expression", *Computer Graphics*, 21, 17-23, 1987.
21. Viaud M.L. & Yahia H. "Facial Animation with wrinkles", *Eurographics'92*, 1992.