SYNTHESIS OF FRICATIVE CONSONANTS BY AUDIOVISUAL-TO-ARTICULATORY INVERSION

Khaled Mawass, Pierre Badin & Gérard Bailly
Institut de la Communication Parlée
46, Av. Félix Viallet, F-38031 Grenoble cedex 01, France
Tel.: +33 (0)4- 76.57.48.26 -- Fax: +33 (0)4- 76.57.47.10, E-mail: mawass@icp.grenet.fr

ABSTRACT

We present here results of audio-visual to articulatory inversion for French fricatives embedded into VCVs. The inversion technique is evaluated using both experimental and synthetic data. The final synthesis is assessed by a perceptual categorisation test. Synthetic stimuli have similar scores as natural ones.

1. INTRODUCTION

Few works [8] have been dedicated to the acoustic-to-articulatory inversion of fricatives. Furthermore the quality of vocal tract analogue synthesis was not sufficient to enable a subjective assessment of the resulting sounds. The feasibility of high quality articulatory synthesis of fricatives has been recently demonstrated at ICP on a limited set of examples [3]. The present paper presents an extension of this work to a larger corpus of Vowel-Fricative-Vowel sequences, an assessment of the inversion procedures used to determine the control parameters of an articulatory synthesiser from audiovisual recordings of a real subject, and a perceptive evaluation of the synthesis.

2. THE ARTICULATORY MODEL AND THE AUDIO-VISUAL DATA

Bergame, the ICP *articulatory synthesiser*, was developed from midsagittal vocal tract contours obtained by cineradiography and recorded in synchrony with front views of the lips for a reference subject [5]. The first module is a physiologically-oriented statistical articulatory model basically driven by nine parameters: *jaw height* JH, *lip height* LH and *protrusion* LP, *tongue advance* TA, *body* TB, *dorsum* TD and *tip* TT, *lip vertical position* LV and *larynx height* LY. The resulting midsagittal vocal tract contours are then converted into area functions, [5] and finally into sound by a simplified aerodynamic model including voice and noise sources. The voice source is controlled by *subglottal pressure* PSG, *rest glottis height* H0 and *vocal fold length* LG, that excites a reflection-type line analogue (more details can be found in [3]).

2.1 LV adaptation

2.1.1 Labiodental constriction and LV

We found that the initial value of the parameter LV that controls the vertical position of the lips with reference to the upper teeth had no influence on the midsagittal and area functions, and thus on the formants. This was particularly unfortunate for the labiodental fricatives, for which the main vocal tract constriction is roughly determined by the lower lip position in relation with the upper incisors. To cope with the *non-audibility* of LV, the labiodental constriction area is now a function of lower lip position.

2.1.2 Acoustics, aerodynamics and LV

Minimal oral constriction area has an obviously important effect on the formants, but is also used to control the fricative noise source (cf. [3]). Because of the short length of the labiodental constriction, the constriction area must be rather small to fit properly the measured formants (typically 0.05 cm^2); on the other hand, aerodynamically equivalent constriction areas for [v] were found to be the order of magnitude of 0.1 cm^2 for the same subject. Therefore, it was decided to use in this case twice the constriction area for the control of the noise source, in order to avoid too small values that tend to stop both voicing and noise generation.

2.2 Corpus

For the present study, a corpus containing the 27 VCV combinations of the French voiced fricative consonants C = [v z ʒ] in all possible vowels contexts with V = [i a u] was uttered by the reference subject on whom the articulatory synthesiser is based. A high quality video system was used to record speech in synchrony with video front views of the subject's lips. The lips were painted blue, in order to facilitate precise lip contours extraction (cf. [7], for a detailed description of the setup).

3. THE AUDIOVISUAL-TO-ARTICULATORY INVERSION

The audiovisual-to-articulatory inversion aims at mimicking the speech produced by a reference subject by determining the appropriate articulatory trajectories of the synthesiser control parameters. Acquiring these parameters by further cineradiography is excluded for obvious ethical reasons. A direct measurement method such as electromagnetic articulometry would be possible, but the setup involved is far from being natural and comfortable for the subject. We have therefore decided to use an indirect estimation by inversion of the articulatory-to-acoustic relation.

The difficulty of acoustic-to-articulatory inversion has been widely discussed (cf. e.g. [1]). The well-known fact that this inversion is an *ill-posed* problem can be overcome by using appropriate *constraints* in the optimisation procedure that derives the articulatory parameters from the acoustic ones.

From the point of view of *robotics*, the articulatory synthesiser can be considered as a *plant.* The *proximal* parameters, i.e. the control parameters of the plant, consist of the nine supralaryngeal articulatory parameters mentioned above. The resulting *distal* parameters are the first four formants F1, …F4, the oral minimum constriction area Ac, and the lip area Al. Note that all the parameters were sampled at the same frequency (100 Hz).

The formants were determined by extracting the roots of LPC coefficient polynomials, carefully correcting them by hand. The intralabial lip area Al was determined from

the video front images of the lips by a pixel counting procedure [7]. As the oral constriction Ac is not directly measurable, it had to be a priori estimated using the time boundaries between the fricative and the adjacent vowels for each VCV sequence. These boundaries were determined as follows. First, the sound power was estimated as a function of time as RMS values calculated over contiguous 10 msec time windows. Three instants were then determined from this curve: the fricative centre, corresponding to the minimum of the power in the sequence, and the centres of the adjacent vowels corresponding to the instants of maximum power on both sides of the fricative centre. Finally, the VC and CV boundaries were determined as the instants when the power reached 40% of the range between the minimum for the fricative and the maxima for the adjacent vowels.

The inversion algorithm is based on a classical gradient descent method: it uses a constrained backpropagation of the configurational error measured between the distal parameters and the target parameters [6]. The algorithm uses a smoothing constraint: the minimisation of the jerk of the proximal parameters. Finally, the error to minimise is the weighted sum of (1) the cumulated quadratic distance between the six targets and the current distal parameters for all the frames in the sequence, and (2) the jerk of the articulatory parameters. More specifically, the formants were converted in barks, in order to give more weight to the lower formants, while the areas were weighted by sigmoids centred in zero so as to give more weight to small constrictions. In fact, the quadratic distances were applied only on each side of *no-error* ranges defined by lower and upper bounds of the distal parameters. The formant ranges were set to ±5% of the target values for all formants; in addition, this range was increased to ±15% for F1 targets below 300 Hz. These choices take into account, among other criteria, the precision of formant measurements. In particular, a high precision measure of low F1's is impossible, because the first or second harmonics of the voice source are mixed with the formant in this frequency region. The lip area range was set to ±10% of the target value. The upper and lower bounds for Ac were determined by interpolating target ranges defined as [0.15 5.0 cm^2] for the vowels (in order to avoid a tendency to closure when F1 is low), and by [0.06 0.1 cm^2] for the fricatives (0.15 cm^2 have been however often found during the inversion of the corpus).

All the articulatory parameters were initialised with zero values, except for LP=3 for [z] fricatives, and for LV=3.5 for [v] fricatives (for reasons discussed below).

The weight of jerk in the minimisation process decreases as a function of number of gradient descent iterations [6]. The temporal smoothing due to the jerk minimisation is most effective at the start of the inversion procedure and vanishes as the process nears the final solution.

In this study, no inversion was performed to determine the three parameters controlling excitation sources. Subglottal pressure PSG was kept constant at 10 cmH_2O; glottal rest height H0 was set to 0.03 cm for the vowels, to 0.035 cm for the voiced fricatives, and interpolated with sigmoids centred on the boundary instants between the fricative and the adjacent vowels; vocal fold length LG was set to 1.6 cm, except for the closed vowels [i u] where is was set to 1.66 cm in order to compensate partly the F0 drop due to the influence of the high vocal tract impedance at low frequencies. It should be added that, in the case of the synthesis of voiceless fricatives, H0 was set to 0.1 cm in the centre of the fricative in order to stop voicing.

4. EVALUATION OF THE INVERSION

The inversion procedure described above is entirely automatic, except for the semi-manual extraction of the formants and for the setting of the initial parameters LP and LV[1], which were chosen depending on the fricative. The inversion was applied to the set of VCVs uttered by the reference subject, and resulted in a corresponding set of synthetic voiced fricative sequences (see an example in Fig. 1). This section presents an objective evaluation of the inversion algorithm.

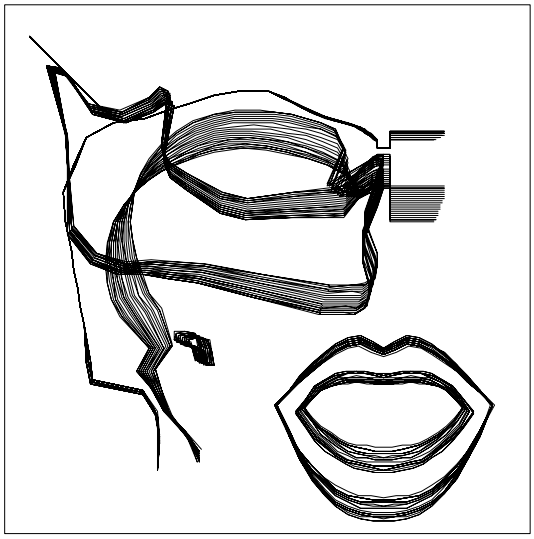

Fig. 1: Evolution of the midsagittal profile for an inverted /aʒa/ .

4.1 Performance on experimental data

The first evaluation can be simply expressed in terms of the residual errors on the distal parameters. Thus, for each sequence, mean quadratic relative errors E_i were computed for each distal parameter (F1..4, Al). Note that no error can be computed for Ac since this parameter was not explicitly measured, but was just given a range of acceptable values depending on the context, as already mentioned above. These errors have been estimated for two conditions: (1) complete *audiovisual inversion*, i.e. using the six distal parameters in the procedure, and (2) *simple audio inversion*, i.e. not taking into account the measured lip area. Table I summarises the results and gives the mean quadratic relative errors computed over the whole set of VCVs. More specifically, it has been found that these errors reach about 12% for F1, and about 5% for F2, F3 and F4, in both inversion cases. The errors on F2, F3, F4 are consistent with the lower and upper bounds used in the computation of the configurational error. The 12% on F1 is also consistent, if we take into account the thresholds applied to the F1 bounds. The error on lip area reaches up to 150 % in some cases of simple audio inversion Fig. 2, but only 40% in the case of audiovisual inversion. Greater discrepancies occur mostly for sequences containing vowel [a] and fricative [v].

[1] Note that a coarse estimation of these parameters could be computed from the speaker's face images.

Error (%)	F1	F2	F3	F4	Al
Audio (±5%)	6.2	2.8	3.0	3.1	56.4
AV (±5%)[2]	6.5	3.4	3.1	3.1	16.8

Table I – Mean quadratic relative errors for natural stimuli (pooled over all VCVs)

Concerning vowel [a], it has already been noticed [2] that the articulatory synthesiser has difficulties in attaining sufficiently high F1's with standard articulations. This is probably due to the parametrisation of the acoustic model and should be solved in the near future. Fig. 2 shows that, for a number of items containing vowel [a] ([ava], [avu], [iva], [aza], [azu], [uza], [aʒa], [aʒu]) the error on lip area was higher for the audio than for the audiovisual inversion, while the opposite occurred for F1 errors. A more detailed analysis has shown that, in the case of audio inversion, the relatively high lip areas (over-estimated by about 100 to 150 %) helped increasing F1 by about 40 Hz, without increasing the configurational cost, since the error at the lips was not taken into account. Inversely, in the case of audiovisual inversion, the constraint on lip area prevented this strategy, thus reducing the error on Al and increasing that on F1.

Concerning fricative [v], it is clear that the lip area can not have any noticeable effect on formants below 5 kHz, since the first resonance associated with the small front cavity shaped by the lips is higher than 7-8 kHz. Thus, the formants of fricative [v] are not expected to be very sensitive to lip area, except for F1 which can decrease with Al (F1 is the Helmholtz resonance between the volume formed by the cavity behind the teeth and the neck formed by the labiodental constriction and the lip horn). This explains for instance the fact that, for [ivi], the small Al error in the audiovisual inversion is not compensated by a greater error on F1.

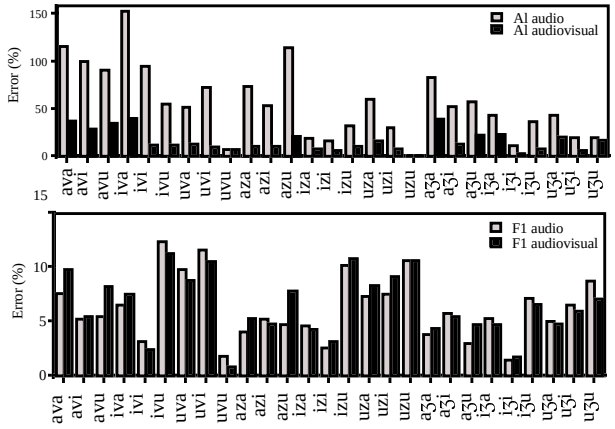

Fig. 2 – Detailed mean quadratic relative errors for lip area (top) and F1 (bottom)

These first results tend to demonstrate that, at least for the present corpus, visual information would not be crucial for the acoustic-to-articulatory inversion. One of the main benefit of lip information is the possibility to infer a better initialisation of LV for [v], and of LP for [ʒ]. Zero initial values for these parameters are actually outside the *audibility* zone of these parameters, i.e. changes of these parameters do not produce any noticeable acoustic changes. We have already mentioned that lip area is not recovered in a number of cases and thus needs additional visual information.

[2] Non error range for Al in the AV condition is ±10%.

4.2 Performance on synthetic data

The precision of inversion from real data, is conditioned by the compatibility between these data and the articulatory-acoustic model. The evaluation described above does not separate out errors due to data and model discrepancies from errors due to intrinsic performances of the inversion algorithm.

Therefore, another evaluation has been carried out with synthetic data. A set of synthetic midsagittal contours, lip areas and formants was derived from the sequences of articulatory parameters found by the first inversion from experimental data, and used as reference data for the evaluation. The audiovisual inversion procedure has been applied to these synthetic references, in two conditions: (1) non-error ranges identical to those in the experimental data inversion; (2) ranges reduced to zero. This led to two new sets of recovered articulatory parameters. Finally, midsagittal contours, lip areas and formants were again computed from these recovered parameters. It was thus possible to assess quantitatively the results of the inversion at both distal (formants and lip areas) and proximal (articulatory parameters and midsagittal distances) levels. Table II summarises the results.

On the average, the quadratic errors on the articulatory parameters obtained for the zero range (0.09) are not significantly lower than those for the ±5% range (0.11). Note that these errors represent roughly 1.6% of the range of normal variation [–3 +3] of the articulatory parameters. The related quadratic relative errors on the midsagittal distances computed over all VCVs pooled together reach about 11% and 6%, for the ±5% and zero ranges respectively; the relative error on the overall vocal tract length VTL (1.2%) is negligible.

The comparison between Table I and II shows that, in the case of equivalent ±5% ranges, configurational errors are 30 to 100% lower for the inversion of the synthetic stimuli than for that of the experimental ones. This difference can be likely ascribed to slight mismatches between the experimental data and the articulatory model.

Range	LH	LP	JH	TB	TD	TT	TA	LY	LV
±5%	0.09	0.10	0.08	0.10	0.18	0.13	0.10	0.13	0.04
0%	0.12	0.09	0.08	0.07	0.13	0.07	0.08	0.12	0.06

Range	F1 (%)	F2 (%)	F3 (%)	F4 (%)	Al (%)	VTL(%)
±5%	3.10	2.60	1.80	1.40	12.6	1.20
0%	1.14	0.47	0.60	0.65	7.00	1.20

Table II – Mean quadratic errors for the nine articulatory parameters and mean quadratic relative errors (in %) for distal parameters with reference to synthetic stimuli (pooled over all VCVs)

5. PERCEPTIVE EVALUATION OF THE SYNTHESIS

As formant extraction for the voiceless fricatives is extremely difficult, not to say impossible, a set of cognate voiceless fricative sequences was derived from the original voiced fricatives, assuming the same

articulatory trajectories; the only difference lay in the glottis opening during the consonant (cf. above).

As one of the main objectives of the present work was to produce articulatory synthesis of fricative consonants, an intelligibility test has been carried out in order to assess the quality of the synthesised sounds. Three sets of stimuli have been therefore tested in a single perception test: (1) voiced fricative VCVs from the original natural stimuli recorded with the audiovisual setup, (2) voiced replications obtained by inversion, and (3) voiceless versions of the same stimuli

Table III gives the exhaustive list of the synthesised sounds that can be played in the CD-ROM version of the proceedings. Ten naive French listeners were thus presented a total of 27*3 = 81 stimuli by means of headphones. They were instructed to classify the embedded consonant of each VCV item as one of the six French fricatives [v z ʒ f s ʃ], with the possibility to replay any item several times. Each stimulus was presented 6 times; the resulting 486 stimuli were randomised, and then divided into four batches of approximately 120 items, so as to allow small rest intervals for the listeners. The typical duration of a complete test was about 40 min.

ava	avi	avu	iva	ivi	ivu	uva	uvi	uvu
aza	azi	azu	iza	izi	izu	uza	uzi	uzu
aʒa	aʒi	aʒu	iʒa	iʒi	iʒu	uʒa	uʒi	uʒu
afa	afi	afu	ifa	ifi	ifu	ufa	ufi	ufu
asa	asi	asu	isa	isi	isu	usa	usi	usu
aʃa	aʃi	aʃu	iʃa	iʃi	iʃu	uʃa	uʃi	uʃu

Table III – List of the VCV synthesised sequences

The results have been analysed and confusion matrices built, the answers of the ten listeners pooled together. Fig. 3 presents these matrices in terms of identification error rates, i.e. ratios of the number of errors for each class over the number stimuli presented, including the rates corresponding to each possible wrong answer. Articulation and voicing errors have been identified and are discussed below for the natural and synthetic stimuli.

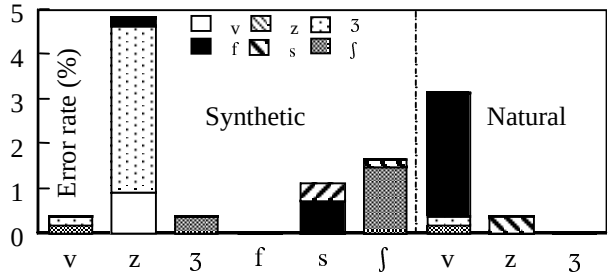

Fig. 3 – Identification error rates (pooled over the nine vocalic contexts)

Natural stimuli

The item [uva] has occasionally been perceived as [ufa] (25% of the cases).

Synthetic stimuli

Consonants [s] and [z] were occasionally wrongly recognised in the context [u-u]: [usu] was recognised as [ufu] in 6.7% of cases, and as [uʃu] in 3.4%, while [uzu] was recognised as [uʒu] in 33.4% of cases and as [uvu] in 8.4%. The analysis of the target articulation of the fricative consonant in this case showed that the constriction was actually made at the lips rather than at the tongue tip, and that the F4 was too high by about 10%.

Beside this articulation problem, a voicing confusion was also found: [iʃu] was perceived [iʒu] in 11.6% of the cases. This is clearly due to a problem of glottis/ constriction coordination (recall that this coordination was achieved in a rather simplified way).

These results show very low identification error rates, and therefore a high intelligibility of the synthesis, except for very few stimuli.

6. CONCLUSION AND PERSPECTIVES

We have shown that audiovisual acoustic-to-articulatory inversion can be successfully carried out on a set of VCV sequences, using a constrained optimisation algorithm based on the gradient descent method. This method present the great advantage to provide useful articulatory data via a parametric estimation. A perceptual test has shown the very good quality of the synthesis: 98.8% recognition rate for 27 natural stimuli versus 98.6% for the 54 synthetic stimuli. In particular, the resulting set of synthesised VCV syllables constitutes the first step towards the establishment of the sensory-motor exemplars needed for a robotic approach of articulatory speech synthesis [4].

ACKNOWLEDGEMENTS

The first author has been granted by the Lebanese Hariri Foundation. We thank C. Vescovi for his advice about the vocal folds two mass model. The software used for perceptive test is a version of the *Europec* software [9], kindly modified by J. Zeiliger, with the help of A. Neagu. We thank also the poor victims of the tests.

REFERENCES

[1] Abry, C., Badin, P., & Scully, C. (1994) Sound-to-gesture inversion in speech: The *Speech Maps* approach. In *Advanced speech applications* (Varghese K. et al., Eds), pp. 182-196. Springer Verlag: Berlin.

[2] Badin P., Gabioud B., Beautemps D., et al. (1995) Cineradiography of VCV sequences: Articulatory-acoustic data for a speech production model. *15th ICA*, Vol. IV, 349-352.

[3] Badin, P., Mawass, K., Bailly, et al. (1996). Articulatory synthesis of fricative consonants: data and models. *4th Speech Production Seminar*, pp. 221-224. Autrans, France.

[4] Bailly, G. (1996) Sensory-motor control of speech movements. 4th Speech Production Seminar, pp. 145-157. Autrans, France.

[5] Beautemps, D., Badin P., Bailly, et al. (1996) Evaluation of an articulatory-acoustic model based on a reference subject. *4th Speech Production Seminar*, pp. 45-48. Autrans, France.

[6] Jordan, M.I. (1990) Motor Learning and the degrees of freedom problem. In M. Jeannerod (Ed.) *Attention and Performance*. Hillsdale, NJ: Lawrence Erlbaum.

[7] Lallouache, M.T. (1990) Un poste "Visage-Parole". Acquisition et traitement de contours labiaux. Actes des 18èmes Journées d'Etude sur la Parole. SFA.

[8] Sorokin, V.N., & Trushkin, A.V. (1996) Articulatory-to-acoustic mapping for inverse problem. Speech Communication 19, 105-118.

[9] Zeiliger, Z., & Sérignat, J.F. (1991) EUROPEC Software, V4.0, User's Guide Release 4.1. Report, ESPRIT Project N°2589 (SAM), Multilingual Speech I/O Assessment Methodology and Standardization, Grenoble.