Synthesis of French Fricatives by Audio-Video to Articulatory Inversion

Khaled Mawass*, Pierre Badin, and Gérard Bailly
Institut de la Communication Parlée UPRESA CNRS 5009, Institut National Polytechnique de Grenoble – Université Stendhal 46, Av. Félix Viallet, F-38051 Grenoble cedex 01, France

Summary
This paper presents an articulatory approach to the synthesis of fricative consonants in vocalic context. The articulatory, aerodynamic and acoustic components of a complete articulatory synthesiser and their interactions are briefly described. The articulatory trajectories of the control parameters for the synthesiser are estimated by inversion from audio-video recordings of the reference subject based on whom the synthesiser was elaborated. The inversion method is described and evaluated. For synthesis, a simple strategy of control of the glottis gesture, in coordination with the oral constriction, was used to produce both voiced and voiceless fricatives. The quality of the acoustic synthesis has been assessed by a forced choice identification test which led to very similar scores for synthetic and natural stimuli. These results validate at the same time the realism of the articulatory synthesiser and the strategies used to control it. Articulatory synthesis appears thus to be able to produce high quality speech sounds, and to provide the ground for the development of a virtual talking head. Moreover, the articulatory data obtained by inversion and the methodology developed will serve as the basis for studying human control strategies for speech production.

PACS no. 43.72.Ja

1. Introduction
Articulatory speech synthesis constitutes one key element for both the study of speech production mechanisms and the development of virtual talking heads. Indeed, articulatory synthesis offers an interesting way to integrate knowledge of speech production in the form of physical and functional models, while allowing the development of audio-visual speech synthesizers capable of producing coherent speech and face animation. To our knowledge, although articulatory synthesis began many years ago (e.g. [1, 2]), the number of fully articulatory synthesizers in use at present is relatively limited [3, 4, 5, 6]. Despite the number of interesting studies devoted to the articulatory modelling of fricative consonants (e.g. [7, 8, 9, 10, 11]), it seems that no systematic work on the articulatory synthesis of vowel-fricative-vowel sequences has been conducted and perceptually assessed so far. A few interesting approaches, though not fully articulatory, should be quoted. Maeda [12], based his approach on the interpolation of area functions between two vowel targets and a consonant target, but did not formally evaluate the results. Stevens and his colleagues [13, 14] also used a composite approach where a formant synthesizer is controlled by parameters derived from a score incorporating acoustic, aeroacoustic and geometric parameters. Finally, the work conducted at Haskins Laboratories should also be mentioned: McGowan et al. [9] have attempted an articulatory approach of the production of syllables (aCa), but in a very limited manner. However, no formal evaluation of the quality of the synthesised sounds, in particular fricative consonants, is available.

For the last few years, we have been developing the different modules of a comprehensive articulatory synthesiser [15, 16, 17, 18, 19]. As a result, the feasibility of high quality articulatory synthesis of fricatives has been recently demonstrated at ICP on a limited set of examples [20]. The present paper describes an extension of this work to a larger corpus of vowel-fricative-vowel (VCV) sequences. A first section offers a brief description of the articulatory synthesiser. A second section presents the audio-video data and the inversion procedures used to determine the control parameters trajectories. The third section is devoted to a formal perceptual evaluation of the synthetic stimuli, and finally some perspectives of this work are discussed.

2. The articulatory synthesiser
As schematised in Figure 1, the ICP articulatory synthesiser consists of a number of interconnected modules, i.e. an articulatory model, an aerodynamic model, acoustic source models and an acoustic vocal tract model.

2.1. The articulatory model
The articulatory model is a physiologically-oriented linear articulatory model [21, 19], developed from midsagittal vocal tract profiles obtained by cineradiography and recorded in synchrony with front views of the lips for a reference subject [22]. This articulatory model allows these vocal tract profiles to be manipulated by a reduced set of nine parameters: jaw height JH, lip height LH and protrusion LP, tongue advance TA, tongue body TB, tongue dorsum TD, tongue tip TT, lip vertical position LV and

Received 3 November 1998, accepted 12 October 1999.
* Present address: Ministry of Telecommunications, Beyrouth, Lebanon
boundary instants between the frictive and the adjacent vowels; vocal fold length LG was set to 1.60 cm, except for the closed vowels [i] and [u] where it was set to 1.66 cm in order to compensate partly for the F0 drop due to the influence of the high vocal tract impedance at low frequencies. For the voiceless fricatives, the supralaryngeal articulators' trajectories were copied from their voiced cognates, while H0 was set to 0.1 cm at the instant of minimum sound power in the fricative segment, in order to ensure that voicing ceased during the consonant. This glottal gesture needs to be carefully coordinated with the trajectory of oral constriction $A_c$ in order to obtain realistic voiceless fricatives [8, 33, 9, 34]. An example of the trajectories resulting is displayed in Figure 5. Note that the volume velocity for the unvoiced case shows the typical double peak for the fricative, but nothing at all similar for the voiced case, though a less well-defined double peak would be expected.

3.4. Evaluation of the articulatory inversion

The inversion procedure described above is entirely automatic, except for the manual correction of the formants, and for the setting of the initial values of LP and LV, which were chosen as a function of the frictive category, but independently of the vocalic substrate (LV was set to 3.5 for [v], and LP to 3.0 for [j]). The inversion was applied to the set of V1CV2 sequences uttered by the reference subject, which resulted, after articulatory inversion from formants and lip aperture and then articulatory synthesis, in a corresponding set of synthetic voiced vowel-fricative-vowel sequences. Before presenting the evaluation of the synthetic speech stimuli themselves in the next section, some results of an evaluation of the inversion algorithm are described.

3.4.1. Distal accuracy

A first evaluation was simply expressed in terms of the residual errors on the distal parameters. Thus, for each sequence, RMS relative errors $E_i$ were computed for each distal parameter (F1, ..., F4, A1). Note that no error can be computed for $A_c$ since this parameter was not explicitly measured, but was just given a range of acceptable values depending on the context, as already mentioned above. Table I contains the resulting RMS relative errors computed over the whole set of VCV sequences. These errors were
found to reach maximum values of about 12% for F1, and of about 5% for F2, F3 and F4. The errors on F2, F3, and F4 are consistent with the limits used in the computation of the configurational error; errors for F1 are similar, if the thresholds applied to the F1 limits are taken into account.

The accuracy of inversion from real data is conditioned by the compatibility between these data and the articulatory-acoustic model. For instance, some measured formant configurations may be difficult if not impossible to reach for the articulatory-acoustic model. In such cases, the inversion would necessarily result into two possible situations: either (1) the formants predicted from the articulatory parameters obtained by inversion from the measured formants are close to these measured formants, but the articulatory parameters obtained by inversion do not match the real articulatory parameters well; or (2) the articulatory parameters obtained by inversion from the measured formants are close to the real articulatory parameters, but the formants predicted from the articulatory parameters obtained by inversion do not match the measured formants well. The evaluation based on real data 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 was carried out with synthetic data.

3.4.3. Importance of lip information

A more thorough study [35] has shown that, at least for the present corpus, the distal parameter $A_1$ is not crucial for the acoustic-to-articulatory inversion. Specifically, Mawass [35] performed two series of inversions on the same corpus of VCV sequences: one with lip area limits set at $\pm 10\%$ of measured $A_1$, and another one with no limits for $A_1$ (actually with a lower limit set to 0.2 cm² in order to avoid lip closure in some cases). The distal errors on formants were very similar for the two experiments, but the lip area was largely over-estimated (up to more than 200%) in the case where $A_1$ was not given an upper limit. The articulatory parameters determined by inversion were also similar in both experiments, except for LH, which is the main determinant of lip area. Some slight articulatory compensations between lip area and tongue shape could be noticed.

However, in some cases, lip rounding information is crucial to avoid certain compensatory effects. Specifically, it was necessary to initialise LV to 3.5 for [v], and LP to 3 for [3] in some contexts. Indeed initialising these parameters to zero values actually push them outside their audibility zone, i.e. into a range of values where varying them do not produce any noticeable acoustic changes, with the consequence that they can not be recovered by inversion from acoustics. This is illustrated in Figure 6, where the

---

1 This was actually done in all vocalic contexts in order to make the procedure more systematic.
inversion results of the sequence [a3a] are shown for two cases. In the case where LP is initialised to zero, some articulatory compensation occurs: the oral constriction is located in the palato-velar region, the lips are relatively rounded and not protruded, the first three formants are in close proximity to the measured formants, but F4 is noticeably too high: this is not a typical [f] phoneme. In the case where LP is initialised with an adequate value of 3, the oral constriction is located in the post-alveolar region, the lips are more open and protruded (which is typical of [f] in French), and all formants are close to the measured ones. It appears thus that lip rounding information is very important in some cases, if not in all.

3.4.4. Importance of specifying limits

The interest of the robotic approach lies in particular in the possibility of specifying limits so that measurement inaccuracies can be accounted for, as already mentioned above about formants.

The interest of specifying limits within the framework of robotics lies also in the possibility of using a distal parameter despite the fact that it cannot be precisely measured. The oral constriction parameter $A_c$ inherently constitutes such a distal parameter, as it is computed by the articulatory model from the control parameters, but cannot be accurately determined from the speech signal. However gross limits can be estimated for it. Preliminary experiments had shown that formants only were not sufficient to ensure a good inversion for vowel-fricative-vowel sequences, and that the oral constriction was often too wide for the fricative. Considering $A_c$ as an extra distal parameter varying within a pair of limits determined from the speech sound power was found useful to solve this problem and to achieve the degree of oral constriction appropriate for fricatives. $A_c$ plays the role of a distal parameter, eventhough it is finally determined as a byproduct of the inversion process, as is lip area $A_l$.

4. Evaluation of the synthetic stimuli

The previous section has described the evaluation of the quality of our articulatory synthesis of fricative consonants at the articulatory level. This section presents an evaluation at the acoustic level, both objectively and perceptually.
Table II. RMS errors for the nine articulatory parameters and RMS relative errors (in %) for distal parameters with reference to synthetic stimuli (pooled over all VCVs).

<table>
<thead>
<tr>
<th>LH</th>
<th>LP</th>
<th>JH</th>
<th>TB</th>
<th>TD</th>
<th>TT</th>
<th>TA</th>
<th>LY</th>
<th>LV</th>
<th>F1 (%)</th>
<th>F2 (%)</th>
<th>F3 (%)</th>
<th>F4 (%)</th>
<th>A1 (%)</th>
<th>VTL (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.09</td>
<td>0.10</td>
<td>0.08</td>
<td>0.10</td>
<td>0.18</td>
<td>0.13</td>
<td>0.10</td>
<td>0.13</td>
<td>0.04</td>
<td>3.10</td>
<td>2.60</td>
<td>1.80</td>
<td>1.40</td>
<td>12.6</td>
<td>1.20</td>
</tr>
</tbody>
</table>

Figure 7. Example of sonograms for the original and synthetic stimuli [a3a] (top: original, bottom: synthesis).

4.1. Spectral evaluation

First of all, the sonograms of the original and synthetic stimuli were visually compared. Figure 7 gives an example of such a comparison. The good quality of the synthetic stimuli can be predicted from the global agreement between original and synthetic sonograms. However, a comparison of spectral sections computed over 100 ms time windows in the centre of the original and synthetic fricative consonants revealed that, for most examples, the amplitudes of formants F2 and F3 were higher for the synthetic stimuli than for the original ones. This formant amplitude discrepancy can very likely be ascribed to an insufficient modelling of the acoustic losses for fricatives. However, it appears that the overall quality of the synthesis is rather high, due to the use of both a realistic articulatory synthesiser and adequate control strategies. This result has been confirmed by a formal perceptual evaluation test.

4.2. Perceptual evaluation

As a crucial complement to the spectral evaluation, an intelligibility test was carried out in order to assess the perceptual quality of the synthesised sounds. Three sets of stimuli have therefore been perceptually tested in a single forced choice test: (1) voiced fricative VCV sequences from the original natural stimuli recorded with the audio-video setup, (2) voiced replications obtained by inversion and articulatory synthesis, and (3) voiceless versions of the same stimuli (all these sounds can be played from the CD-ROM provided with the Journal, see the lists in Tables IV–VI).

Ten naive French listeners were thus presented a total of 27*3 = 81 stimuli by means of headphones. They were instructed to label the embedded consonant of each VCV item as one of the six French fricatives [v z ʒ f s j], with the possibility of replaying any item any number of times. Each stimulus was presented 6 times; the resulting 486 stimuli were randomised, and divided into four batches of approximately 120 items, so as to allow small rest intervals for the listeners. The typical duration of the whole test was about 40 minutes.

The overall identification rates were nearly identical for the original and synthetic stimuli (98.8% for the original stimuli, and 98.6% for the synthetic ones). These rather good results were expected, as [s, j] are the most easily distinguished fricatives in any language, and as [f] is also always well distinguished from [s, j]. Moreover, French is a language in which devoicing of voiced fricatives occurs mainly in consonantal context and never occurs in intervocalic position. Thus the voiced–voiceless distinction here is likely to be easy. The present results, that demonstrate the rather high intelligibility of the present articulatory synthesis, should however be further confirmed by more challenging experiments such as adding background noise at different signal to noise ratios in order to compare the robustness of the intelligibility of the synthetic sounds against that of the natural ones.
larynx height LY [21, 19]. Note that all these parameters are normalised, and evolve within \([-3 \pm 3\) for speech. The resulting midsagittal vocal tract profiles are then converted into area functions by a conversion algorithm optimised on the same subject’s data [21, 19]. The effects of the control parameters on the midsagittal profiles are exemplified in the nomograms presented in Figure 2.

2.1.1. Specific lip vertical movement

A careful analysis of lip geometry parameters has revealed three main degrees of freedom, in addition to LH: a factor related to lip height, the distance between upper and lower lips, taken into account by LH, a factor related to both upper and lower lip protrusions, LP, and a third factor, related to a mere vertical synchronous movement of both upper and lower lips relative to the lower edge of upper incisors, taken into account by the lip vertical position parameter LV. Parameters LP and LH present obvious acoustic correlates, as they determine the size of the final vocal tract section, but LV has in principle no influence on the area functions, and thus on the formants. Neglecting the teeth in the model was particularly problematic for the labio-dental fricatives, for which the main vocal tract constriction is roughly determined by the lower lip position in relation with the upper incisors. In order to use this parameter in the production of labio-dental fricatives, the labio-dental constriction was realised at the section of the incisors; its area was limited by a minimum threshold dependent of lower lip position (see Figure 3) and thus indi-

rectly made a function of LV. This allows the LV articulatory parameter to be audible, i.e. to have acoustic consequences, at least in circumstances typical of labio-dentals where the lower lip has to be higher than the upper incisor edge in order to produce the proper constriction. This feature is very useful for the inversion of the articulatory-to-acoustic relation.

2.2. The low frequency airflow model

The aerodynamic phenomena that convert the lung pressure into acoustic excitation sources for the vocal tract are taken into account by a simplified airflow model, valid at low frequencies (below approximately 200 Hzs). This model considers the vocal tract as a set of two constrictions: the glottis \(A_g\) and the oral constriction \(A_c\). Bernoulli and Poiseuille equations are used to express the pressure drop across the constriction \(\Delta P_c\), as a function of \(A_c\), and the pressure drop \(\Delta P_g\) across the glottis as a function of \(A_g\), where \(A_g\) is the low frequency component of the glottal area. In this low frequency lumped constriction approximation, the sum of \(\Delta P_g\) and \(\Delta P_c\) is equal to the subglottal pressure PS, which is one of the command parameters of the articulatory synthesiser. This airflow model is essential for the coordination and the interactions between the vocal tract geometry and acoustic sources. This airflow
model is essential for ensuring: (1) the proper influence of the voice source upon the noise source (the modulation of the noise source amplitude induced by the vocal tract flow variations generated by the voice source), and (2) the influence of vocal tract oral constriction upon voicing (the reduction of voicing amplitude due to the increase of pressure drop at the oral constriction and the concomitant decrease of pressure drop at the glottis).

The voice source is a two-mass model of the vocal folds [23, 24] controlled by trans-glottal pressure $\Delta P_g$, rest glottis height $H_0$ (i.e. the distance between vocal folds) and vocal fold length $L_G$. This model delivers the flow at the glottis, the derivative of which is used as the source of voicing in the acoustic model. The resulting glottal area signal is low-pass filtered in order to retain only the slow variations at the fundamental frequency and is re-injected in the airflow model (cf. Figure 1). Note that the coefficients of this model have been tuned in order to be able to reproduce as faithfully as possible the reference subject’s voice [16].

The friction noise is produced by a functional model that predicts noise spectral characteristics from cross-sectional area and pressure drop at the oral constriction [18]. This model is controlled by the low frequency component of the pressure drop $\Delta P_c$ across the oral constriction and by the main constriction area. This ensures that the low frequency flow fluctuations at the glottis are transmitted to the constriction pressure drop and used to modulate synchronously the friction source in the case of voiced fricatives (cf. [25]). We assume, for the three classes of French fricatives, labio-dentals [f v], predorsal-alveolars [s z], and post-alveolars [s z], that the main source of friction noise is created at the upper incisors by the air jet hitting them [26, 27]. The source is thus considered as a pressure source inserted in series between the upper incisors section and the one immediately downstream; this assumption has been validated by frequency domain simulations [26].

2.3. The acoustic model

The resulting synthetic speech sound is produced by a time-domain reflection-type line analogue [15], that takes into account the acoustic interaction between the vocal tract acoustic model and the two-mass model.

The complete articulatory synthesiser is thus globally controlled by two sets of articulatory parameters: supralaryngeal parameters (i.e. the command parameters of the articulatory model), and laryngeal parameters controlling the vocal folds (glottal pressure $P$, vocal folds length $L_G$, glottis rest height $H_0$), that need to be carefully coordinated, as will be discussed further in this paper.

3. Synthesis strategy

Because one of the major goals of this study was to assess the possibilities of achieving high quality articulatory synthesis for some French fricative consonants, and also because the different modules of the articulatory synthesiser were based on the same subject, it was decided to perform copy synthesis by mimicking the speech produced by this reference subject. However, acquiring the appropriate articulatory trajectories of the synthesiser control parameters by further cineradiography was excluded for obvious ethical reasons. Direct measurement methods such as electromagnetic articulometry would have been possible, but the setup involved is far from being natural and comfortable for the subject. An indirect estimation of these articulatory parameter based on the inversion of the articulatory-to-audio-visual relation was therefore chosen.

3.1. Reference audio-video data

The reference subject was thus asked to utter the set of 27 $V_1CV_2$ combinations of the French voiced fricative consonants $F \in \{v z \}$ in all possible vowels contexts with $V_1$ and $V_2 \in \{i a u \}$. Since formants in voiced fricatives can be easily identified and detected, whereas their extraction in voiceless fricatives is extremely difficult, voiced fricatives only were recorded. A high quality video system (a 3CCD camera JVC KY-15E and a UMATIC-SP video recorder) 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. [28] or [29] for a detailed description of the setup). The temporal trajectories of four acoustic and two geometric parameters were estimated from these recordings: the first four formants $F_1$ to $F_4$ were determined by extracting the roots of LPC polynomials, carefully corrected by hand, while intra-labial lip area $A_l$ was determined from the internal lip contours automatically extracted from the lip images [28]. As the oral constriction area $A_c$ cannot be directly measured from these recordings, it was merely constrained to vary between arbitrary limits in the vocalic and fricative segments of each VCV sequence.

These limits were determined from the speech sound power, based on the idea that the sound power varies in a roughly monotonic way with the oral constriction area.
\(A_v\), i.e. the sound power decreases from open vocalic articulations to constricted fricative articulations, and vice-versa. The boundaries between vocalic and fricative segments were determined as follows. First, the sound power was estimated from the speech sound as a function of time as RMS values calculated over contiguous 10 msec time windows. Three instants were determined from this curve: the instant of minimum power in the sequence, roughly corresponding to the fricative centre, and the instants of maximum power on both sides of the fricative segment, roughly corresponding to the centres of the adjacent vowels. Finally, VC and CV boundaries were taken as the instants when the power reached 40% of the range between the minimum for the fricative and the maxima for the adjacent vowels, and the values of \(A_v\) limits were interpolated by means of sigmoidal functions between arbitrary values (see further).

3.2. Articulatory inversion from formants and constrictions

The difficulty of pure acoustic-to-articulatory inversion has been widely discussed (cf. e.g. [30]). The well-known fact that this inversion is an ill-posed problem, because of the many-to-one nature of the articulatory-to-acoustic/geometric relation, 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 plant is controlled by proximal parameters and produces distal parameters [31, 32]. The proximal-to-distal transform is usually a many-to-one function, and thus the inverse transform cannot be explicitly defined. In the present case, the proximal parameters, i.e. the control parameters of the articulatory model, consist of the nine supralaryngeal articulatory parameters mentioned above. The resulting distal parameters consist of four acoustic parameters, the formants F1 to F4, and of two geometric parameters, the minimum oral constriction area \(A_v\) and the lip area \(A_l\). The inversion algorithm, in the framework of a robotic approach, aims at overcoming the impossibility of deriving directly the articulatory parameters from the acoustic and geometric ones. Note that all the parameters are sampled at the same frequency (100 Hz).

The inversion algorithm is based on a classical constrained gradient descent method: it uses the back propagation of the configurational error between the measured and computed distal parameters [32]. The algorithm also uses a smoothness constraint, the minimisation of the acceleration of the proximal parameters. Finally, the error minimised is the weighted sum of: (1) the quadratic distance between the six measured acoustic and geometric parameters and the six computed distal parameters accumulated over all the frames in the sequence, and (2) the cumulated acceleration of the articulatory parameters. 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 on zero so as to give more weight to small constrictions.

3.2.1. Distal lower and upper limits

In fact, each distal parameter was allowed to move freely within predefined lower and upper limits, within which the distal error is set to zero; the quadratic distances are then computed between the distal parameter and the upper limit if the parameter lies above the upper limit, and similarly between the distal parameter and the lower limit if the parameter lies below the lower limit. For the fricative segments, the upper and lower limits for the formants were set to \(\pm 5\%\) of the measured values for all formants; for the vocalic segments, this range was set to zero. In addition, this error free range was increased to \(\pm 15\%\) for F1 measured values 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 limits was set to \(\pm 10\%\) of the measured \(A_l\) value. The values of the upper and lower limits for \(A_v\) were taken as \([0.15-6.0]\) cm² for the vowels; the lower value is chosen in order to avoid complete closure of the constriction when F1 is low, which could occur either due to a too low measured value of F1 or to a too high value of the close lip vocal tract resonance used to represent the wall vibrations in the acoustic model. These limits were set to \([0.06-0.1]\) cm² for the fricatives.

As mentioned above, sigmoidal interpolations between the centers of vocalic and fricative segments were used to establish the complete trajectories of the limits.

3.2.2. Initial values for proximal parameters

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 articulators acceleration in the minimisation process decreases as a function of number of gradient descent iterations [32]. The temporal smoothing due to the acceleration minimisation is most effective at the start of the inversion procedure and vanishes as the process approaches the final solution.

Figure 4 illustrates inversion results for the sequence [a3a]. The trajectories of the upper and lower limits for the distal parameters are displayed, as well as those of the same parameters predicted from the proximal articulatory parameters obtained by inversion. The trajectories of these articulatory parameters are also shown, as is the corresponding evolution of the vocal tract profile.

3.3. Control strategies for excitation sources

In this study, no inversion was performed to determine the three parameters controlling acoustic excitation sources, PS, H0 and LG (the parameter \(A_v\) is determined as a byproduct of the inversion process, cf. infra 3.4.4). The same strategy was used for all the voiced fricative sequences: subglottal pressure PS was kept at a constant value of 10 cm H₂O; 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
Figure 8 presents more detailed results in terms of identification error rates, i.e. ratios of the number of errors for each class over the number of stimuli presented, including the details corresponding to each possible erroneous answer. Table III provides the complete confusion matrix in a simplified form. Some errors that occurred once or twice only (for a total of sixty presentations for the whole set of subjects) can be very likely attributed to response typing mistakes, as some subjects reported, and will thus be excluded from the discussion. Four significant errors remain.

The original item [uva] was occasionally perceived as [ufa] (25% of the cases), due to a slight mispronunciation of the reference subject; interestingly the synthetic versions [uva] and [ufa] were correctly identified!

The synthetic consonants [s] and [z] were occasionally wrongly identified in the context [u–u]: [usu] was identified as [usu] in 6.7% of cases, and as [usu] in 3.4%, while [usu] was perceived as [usu] in 33.4% of cases and as [usu] in 8.4%. The analysis of the target articulation of the fricative consonant in this case showed that the constriction determined by the inversion was actually realised with the lips rather than with the tongue tip, and that the F4 was too high by about 10%.

Besides this articulation problem, a voicing confusion was also found: [ufi] was perceived of [ufi] in 11.6% of the cases. This was clearly due to a problem of glottis / constriction coordination (recall that this coordination was not achieved in a very detailed way).

The listener responses indicated that the [a] vowels are sometimes a little centralised, their F1 being too low. It has already been noticed [22, 19] that the articulatory synthesiser encounters some difficulties in attaining sufficiently high F1's with standard articulations. This can likely be ascribed to the method of obtaining the area function from the sagittal profile in the vicinity of the epiglottis, and should be solved in the future by three-dimensional modelling of the vocal tract.

Concerning the prosody of the synthesised items, the listener can easily notice that F0 tends to be lower in vowels [i] and [u]. This is primarily due to the aerodynamic coupling between constriction area and two-mass model. Since the focus of this study was on the fricatives, we did not attempt to compensate systematically for these interactions by adjusting the vocal fold length (a slight adjustment was made for vowels [i] and [u], cf. above).

4.3. Voicing / frication noise ratio in voiced fricatives

The adequate balance between the contribution of voice and frication acoustic sources during the fricative consonants plays an important role in the quality and naturalness of the synthesis (cf. e.g. [8]). We have recently shown, using the airflow model described in the first section, that the adequate voicing / frication noise balance depends on a critical coordination between glottis and oral constriction coordination [35, 34]. In the present vocal tract model, the frication noise source is controlled by a variable gain that depends on $\Delta P_c$ and $A_c$ as mentioned above, applied to a noise source of constant arbitrary amplitude; during the phase of the synthesiser design, this constant amplitude was set so as to yield, on average, the same proportion of voice and frication components in the synthetic voiced fricatives as in the natural ones used to establish the frication source parameters (cf. [18]). In order to test the relevance of this choice at the perceptual level, a small informal test has been carried out. For a reduced set of VCV sequences ([ava], [aza], [aza]), the amplitude of the noise source was varied by 6dB steps between −24dB and +24dB around its default value, leading to nine versions of each stimuli. An informal listening test confirmed that the preferred level lies between −6 and +6dB, as the reader can test for him/herself by listening to these sounds on the CD-ROM. Note that in the case of [ava], the perception tends towards [ava] when the noise level is too low.

5. Conclusion

We have demonstrated the possibility of performing high quality articulatory synthesis of vowel-fricative-vowel sequences in French. Articulatory inversion from formants and lip aperture using a constrained optimisation algorithm based on the gradient descent method was successfully employed to determine the articulatory control parameter trajectories. A simple strategy of control of the glottis gesture, in coordination with the oral constriction, was used to produce both voiced and voiceless fricatives. A formal perceptual test has demonstrated the rather high quality of the synthesis: 98.8% identification rate for 27 natural stimuli versus 98.6% for the 54 synthetic stimuli. These results validate at the same time the realism of the articulatory synthesiser and the strategies used to control it.
Table IV. List of the sound samples. 1. Original sounds, followed by the voiced and voiceless versions of the synthesis.

<table>
<thead>
<tr>
<th>Original</th>
<th>Synthesis (voiced)</th>
<th>Synthesis (voiceless)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ava</td>
<td>ava</td>
<td>afa</td>
</tr>
<tr>
<td>avi</td>
<td>avi</td>
<td>afi</td>
</tr>
<tr>
<td>avu</td>
<td>avu</td>
<td>afu</td>
</tr>
<tr>
<td>iva</td>
<td>iva</td>
<td>ifa</td>
</tr>
<tr>
<td>ivi</td>
<td>ivi</td>
<td>ifi</td>
</tr>
<tr>
<td>ivu</td>
<td>ivu</td>
<td>ifu</td>
</tr>
<tr>
<td>uva</td>
<td>uva</td>
<td>ufa</td>
</tr>
<tr>
<td>uvi</td>
<td>uvi</td>
<td>ufi</td>
</tr>
<tr>
<td>uvu</td>
<td>uvu</td>
<td>ufu</td>
</tr>
<tr>
<td>aza</td>
<td>aza</td>
<td>aza</td>
</tr>
<tr>
<td>azi</td>
<td>azi</td>
<td>azi</td>
</tr>
<tr>
<td>azu</td>
<td>azu</td>
<td>asu</td>
</tr>
<tr>
<td>iza</td>
<td>iza</td>
<td>isa</td>
</tr>
<tr>
<td>izi</td>
<td>izi</td>
<td>isi</td>
</tr>
<tr>
<td>izu</td>
<td>izu</td>
<td>isu</td>
</tr>
<tr>
<td>uza</td>
<td>uza</td>
<td>usa</td>
</tr>
<tr>
<td>uzii</td>
<td>uzii</td>
<td>usii</td>
</tr>
<tr>
<td>uzii</td>
<td>uzii</td>
<td>usii</td>
</tr>
<tr>
<td>aza</td>
<td>aza</td>
<td>aza</td>
</tr>
<tr>
<td>azi</td>
<td>aza</td>
<td>azi</td>
</tr>
<tr>
<td>azu</td>
<td>azu</td>
<td>azu</td>
</tr>
<tr>
<td>iza</td>
<td>iza</td>
<td>iza</td>
</tr>
<tr>
<td>izi</td>
<td>izi</td>
<td>izu</td>
</tr>
<tr>
<td>izii</td>
<td>izii</td>
<td>izii</td>
</tr>
<tr>
<td>izii</td>
<td>izii</td>
<td>izii</td>
</tr>
<tr>
<td>uza</td>
<td>uza</td>
<td>uza</td>
</tr>
<tr>
<td>uzii</td>
<td>uzii</td>
<td>uzii</td>
</tr>
<tr>
<td>uzii</td>
<td>uzii</td>
<td>uzii</td>
</tr>
</tbody>
</table>

Table V. List of the sound samples. 2. Variations of friction noise source gain between -24 and +24 dB.

<table>
<thead>
<tr>
<th>Original</th>
<th>Synthesis (voiced)</th>
<th>Synthesis (voiceless)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ava</td>
<td>(-24, -18, -12, -6, 0, +6, +12, +18, +24)</td>
<td></td>
</tr>
<tr>
<td>aza</td>
<td>(28-a, 28-b, 28-c, 28-d, 28-e, 28-f, 28-g, 28-h, 28-i)</td>
<td></td>
</tr>
<tr>
<td>aza</td>
<td>(-24, -18, -12, -6, 0, +6, +12, +18, +24)</td>
<td></td>
</tr>
<tr>
<td>aza</td>
<td>(29-a, 29-b, 29-c, 29-d, 29-e, 29-f, 29-g, 29-h, 29-i)</td>
<td></td>
</tr>
<tr>
<td>aza</td>
<td>(-24, -18, -12, -6, 0, +6, +12, +18, +24)</td>
<td></td>
</tr>
</tbody>
</table>

Table VI. List of the sound samples. 3. Synthesis of the sentence:

"Sophie, je suis fâché! Vous savez?":

In order to evaluate the developments that can be envisaged from this study for more continuous speech, a complete fricative-rich sentence has been synthesized: "Sophie, je suis fâché, vous savez?" ("Sophie, I am angry! You know?"). The quality of the synthesis demonstrated on the CD-ROM confirms the interest of the articulatory approach. In order to assess more thoroughly the quality of the synthesis, and in particular its robustness to adverse conditions, a series of identification tests with different signal to noise ratios is planned.

The development of talking heads (cf. [36]) for audiovisual speech synthesis (cf. e.g. [37] or [38]) or for aids to pronunciation training for foreign languages (cf. e.g. [39]) will benefit from the present advance in the articulatory synthesis of fricative consonants. Moreover, the articulatory data obtained by inversion and the methodology developed will serve as the basis for studying human control strategies for speech production. In particular, the resulting set of synthesised VCV sequences constitutes the first step towards the establishment of the sensory-motor exemplars needed for a robotic approach to articulatory speech synthesis [40].

6. Sound illustrations

The collection of sounds illustrating this study and referenced in Tables IV, V and VI will be published in a forthcoming CD-ROM by Acustica / acta acustica. A version of them is readily available at the following web address: http://www.icp.inpg.fr/badin/AcutaAcustica_Sounds.html

Acknowledgement

This work has been partially funded by the European collaborative project Speech Maps, and by a grant of the Lebanese Hariri Foundation to the first author. The data forming the basis of this work owe a lot to many colleagues who collaborated on Speech Maps, in particular B. Gabioud, C. Scully and C. H. Shadle. We also thank X. Pelorson and C. Vescovi for their advice about the twomass model of the vocal folds. The software used for the perceptual test is a version of the Europec software [41] kindly modified by J. Zeiligier, with the help of A. Neagu. We thank also the poor victims of the tests! Finally, we are very grateful to two anonymous reviewers for their pertinent comments and careful editorial advice.

References


