Auditory identification and acoustic representation of the voiceless fricatives and affricates

Sergio Feijóo and Santiago Fernández

Departamento de Física Aplicada, Universidad de Santiago de Compostela

Abstract

The voiceless fricatives and affricates of Galician, /f/, /θ/, /s/, /ʃ/, /x/ and /tʃ/, have been studied from an acoustical and perceptual point of view in order to find the relation between a dynamic spectral representation based on FFT derived linear cepstral coefficients and the auditory identification. Two different conditions were investigated: isolated fricative noise (F condition), and fricative noise plus 51.2 ms of the following vowel (FV condition). The acoustic distance from an individual token to a certain group was defined as the *a posteriori probability* of group membership of the token in that group. The listeners' responses were characterized by *response profiles*. The correlations between acoustic distances and perceptual distances were used as indicators of the match between the acoustic and the auditory properties of the tokens. The results show that voiceless fricatives and affricates are better defined acoustically and perceptually in the FV condition. A low order spectral analysis (4 coefficients) yielded the best results. Nevertheless, the acoustic method did not replicate the perceptual results equally well across all phonemes, indicating that the perceptual identification of voiceless fricatives and affricates is not entirely based on the dynamic evaluation of global spectral shape.

1 Introduction

The purpose of the present paper is to study the relation between the spectral representation and the auditory identification of the voiceless fricatives and affricates of Galician: /f/ (labiodental), /θ/ (linguo-dental), /s/ (alveolar), /ʃ/ (palato-alveolar), /x/ (velar), and the affricate /tʃ/ (palato-alveolar). Knowledge of that relationship would permit us to improve the performance of automatic speech recognition systems (ASR), particularly those dealing with syllables (for instance speech learning applications for hearing impaired people). Since fricatives and affricates share many common characteristics, an ASR system may have difficulties distinguishing between the two types of sounds.

Fricative consonants are produced when the air flow coming from the lungs passes through a constriction in some point of the vocal tract, creating a turbulent flow. Affricate consonants are dynamical sounds that can be considered as the combination of a stop and a fricative, usually with different places of articulation (Strevens, 1960). For instance, the voiceless affricate /tʃ/ can be modelled as the combination of the stop /t/ with the fricative /ʃ/ (Rabiner and Schafer, 1978), the release burst having a spectrum closely similar to that of /s/ (Strevens, 1960). Acoustically, the main difference between /ʃ/ and /tʃ/ is the presence of a stop-like release preceding the fricative noise (Olive et al., 1993). Nevertheless, the release mechanism of the affricate consonant is not entirely similar to that of a stop consonant (Stevens, 1993). The initial part of the release is influenced by the shape of the cavity that is anterior to the closure formed by the front part of the constriction. Approximately 50 ms after the release, the mechanical properties of the affricate are similar to those of a fricative consonant.

Auditory identification of consonants depends on the interaction between the acoustic properties of the consonant and the accompanying vowel, since adjacent segments in the speech signal are not independent in the auditory system. This percep-

e-rthabla, N1, Abril 2003, ISSN: 1695-9914

tual integration often seems to be an obligatory task over which little control can be exerted by the listener (Nygaard and Pisoni, 1995; Repp, 1988). Numerous perceptual studies have shown that although invariant cues are available in the friction noise, at least for some fricatives, listeners also rely on context-dependent cues (see for instance Whalen (1991); Borzone de Manrique and Massone (1981)). In particular, the distinction between /θ/ and /f/ depends on cues contained in the vocalic part (Harris, 1958; La Riviere et al., 1975; Jongman, 1989). In their study about fricative perception by normal-hearing and hearing-impaired listeners, Zeng and Turner (1990) found that normal listeners take advantage of the relevant characteristics of both the friction noise and of the vocalic transitions. The difficulty for hearing-impaired listeners was the lack of audibility of the relevant characteristics of the friction and their inability to use the information contained in the vocalic transitions as efficiently as normal listeners do. The perceptual identification of the fricative depends on the compatibility between the spectral characteristics of the fricative noise and the vowel (Behrens and Blumstein, 1988).

Most of the studies on the automatic classification of fricatives have been based on parameters describing the global spectral shape of the frication noise. Some of the problems in the characterization of fricatives have been described by Shadle and Mair (1996). Basically, the influence of vowel context, speaker's vocal tract length and other factors tend to affect the performance of parameters based on detailed spectral cues. Therefore, parameters based on global spectral cues are preferred, particularly statistical moments, which have been used in several works (Shadle and Mair, 1996; Forrest et al., 1988; Jongman et al., 2000; Flipsen et al., 1999; Jassem, 1979). Although spectral measures based on the LPC spectrum have been used by some authors (Jassem, 1998; Feijóo et al., 1999), the inability of the LPC method to model zeroes (related to the presence of troughs in the spectrum) makes their use questionable. Some of the latter studies also investigated whether distinctions in terms of place of articulation are more successfully captured by static or dynamic properties of the signal. To date, no study has systematically investigated the effect of including the vowel on the perceptual identification of the fricative and its acoustical consequences.

In this paper fricative-vowel syllables in word initial position are analysed from an acoustical and perceptual point of view. Since the frication noise and the vowel contribute to the perception of the fricative, both of them have been used to characterize the fricative. A dynamical description of the global spectral shape of both the frication noise and the accompanying vowel was obtained computing the trajectories of FFT derived cepstral coefficients with different orders. The acoustic distance from an individual token to a certain group was defined as the *a posteriori probability* of membership in each reference group (APP). The listeners' responses were characterized through *response profiles*. Finally the correlation between listeners' responses and the acoustic distances was calculated in two conditions: a) Isolated fricative noise; and b) Fricative noise plus 51.2 ms of the following vowel. The first objective of our work is to assess for which of the two conditions the correlation between the observed and predicted responses is higher. If both frication noise and vowel contribute to the fricative perception, the fricative should be better defined in the fricative + vowel condition, and the correlation with the acoustic distances should be higher for that condition, since confusing stimuli, not being univocally defined, may cause a decrease in correlation. The second objective is to determine for which order the best match between the spectral representation and the auditory identification is obtained in both the F and FV conditions.

2 Perceptual experiments

2.1 Stimuli

The corpus of utterances consisted of 30 two-syllable natural words of Galician (FVCV). The voiceless fricatives /f, θ, s, ʃ, x/ and the voiceless affricate /tʃ/, were followed by the vowels /a, e, i, o, u/ in initial syllable position. All the words were stressed in the first syllable. Twenty subjects, ten men and ten women, with ages between 20 and 30 years old, served as speakers. All of them were native speakers of Galician with no known history of speech or hearing disorders. Therefore the total number of stimuli was $600 = 6$ fricatives $\times$ 5 vowels $\times$ 20 speakers.

The thirty words were uttered by each speaker

e-rthabla, N1, Abril 2003, ISSN: 1695-9914

in a normal office in the Faculty of Physics. The subjects were asked to pronounce the words in a natural way, the distance between the microphone and the lips being chosen so as to take advantage of the whole quantization range. No further instructions were given to the talkers. The signals were captured with a microphone (Rion, type UC-53A), digitized with a 20 kHz sampling rate and a precision of 12 bits/sample, and stored on the computer disk. The signals were filtered using a Chebyshev band-pass filter with cutoff frequencies of 100 and 9200 Hz.

Since the words in our corpus were produced at different effort levels, depending on the particular speaking style of the speaker, the amplitude of the fricative noise varied a great deal among the different speakers for a given fricative. Therefore, the amplitude of the signals was normalized with respect to the vowel. A vowel intensity normalization similar to that of Zeng and Turner (1990) was used. The maximum amplitude value of each signal was determined, and all the samples of the signal were normalized with respect to that maximum value. Since the words were stressed on the first syllable, the maximum amplitude value corresponded to the vowel following the fricative.

The signals were manually segmented. The first segment corresponded to the fricative noise, from the start until the beginning of vowel onset. The point of vowel onset was located at the beginning of the first clear pulse of the vowel. Sometimes that point was hard to find, due to the presence of fricative noise mixed with the voicing pulse. In those cases the LPC spectrum of the pulse was checked looking for a steep rise of the second formant. The cutting points are automatically relocated by the program at the closest zero-crossing point. Two segments were selected for each signal: the first one is the fricative noise, henceforth denoted as F, and the second is the fricative noise plus 51.2 ms of the vowel from vowel onset (which is the end of the fricative), henceforth denoted as FV. Preliminary work showed that the coarticulatory effects due to the fricative are restricted to the first 10–40 ms of the vowel (Feijóo and Fernández, 2002), so the stimuli contained all the information related to the fricative's place of articulation. The initial part of the fricative and the final part of the vowel were smoothed using a cosine type, 10 ms long window, in order to avoid the presence of clicks due to a sudden rise or fall of the amplitude. This smoothing did not affect the auditory quality of the phoneme /tʃ/. Since the duration of the fricative noise varied with each signal, the duration of the F and FV stimuli were not kept constant across the signals.

Two different experiments were performed based on two signal parts: a) Isolated fricative noise (F condition) and b) Fricative noise plus 51.2 ms of the following vowel (FV condition).

2.2 Subjects

Eleven native speakers of Galician with ages between 20–25 years participated as listeners in the experiments. None of them participated as talker in the recording of the signals. Nine were phonetically naive students and two were students helping with the experiments. They participated voluntarily in the experiments without being paid and, prior to the experiments, all of them passed and audiometric test.

2.3 Procedure

The experiments took place in a quiet office of the Faculty of Physics. The two experiments, performed by each listener in different days due to the size of the sample, were controlled by a perceptual analysis program developed at our laboratory. The signals were presented binaurally through Sony MDR-CD5790 headphones at an approximate A-weighted sound level of 70 dB, in the FV condition. In the F condition, the amplitude of the fricative noise was the same as in the FV condition. The experiment was self-paced and the random order of presentation was different for each listener. In the experiment only one repetition of the signal was allowed, after which it was mandatory to select one answer. All the procedures and characteristics of the perceptual analysis program were written in a small script and explained orally to the subjects.

In both experiments the task for the listeners was the same: to identify the consonant in initial position as one of the fricatives /f, θ, s, ʃ, x/ or the voiceless affricate /tʃ/. In order to reduce the amount of guessing by the listeners an extra option was included (i.e., *another sound*).

e-rthabla, N1, Abril 2003, ISSN: 1695-9914

2.4 Results and discussion

Identification rates were 81.8% in the F condition and 88.0% in the FV condition.

Table 1 (top) shows the confusion matrix for the F condition. The fricative /θ/ has the lowest identification rate (56.5%), while /x/ and /tʃ/ are recognized from the isolated fricative noise almost perfectly (99.5% and 97.3%, respectively). The responses of the listeners in that experiment were submitted to a three-way analysis of variance (fricative × vowel × sex). There was only a significant main effect for fricative ($F(5, 594)=66.099$, $p<0.0005$), plus a significant fricative × vowel interaction ($F(20, 579)=2.071$, $p<0.004$). The same data was submitted to a one-way analysis of variance and the Scheffé test (significance $p<0.05$) with the fricative as the independent factor, in order to find the groupings of fricatives ($F(5, 594)=64.304$, $p<0.00005$). The Scheffé test divided the data in the following groups: Subgroup 1 (/θ/); Subgroup 2 (/f, ʃ, s/); Subgroup 3 (/tʃ, x/). Then the results indicate that /θ/ is the phoneme with the lowest score, followed by /f, ʃ, s/, which attain a similar identification rate, while /x/ and /tʃ/ are better identified than the rest in the isolated fricative noise condition. To gain an insight into the fricative × vowel interaction, each fricative was considered separately, and the responses were submitted to a one-way analysis of variance and the Scheffé test (significance $p<0.05$), with the vowel as the independent factor. The effect of the vowel was only significant for the fricative /f/ ($F(4, 95)=7.534$, $p<0.0005$), with the following subgroups: Subgroup 1 (/e, a, i/); Subgroup 2 (/a, i, o/); Subgroup 3 (/i, o, u/): /f/ is recognized better in the context of /u/, while the worst recognition occurs in the context of /e/. Figure 1 (top) shows the correct recognition scores of the phonemes in each vowel context. A similar one-way analysis of variance was performed on the same data with the sex of the speaker as the independent factor. No sex related effects showed up for any of the fricatives.

Table 1 (bottom) shows the confusion matrix for the FV condition. /x/ and /tʃ/ are again identified almost perfectly (100% and 98.5%, respectively), while differences among the rest of the fricatives are smaller than in the F condition. The same three-way analysis of variance (fricative × vowel × sex)

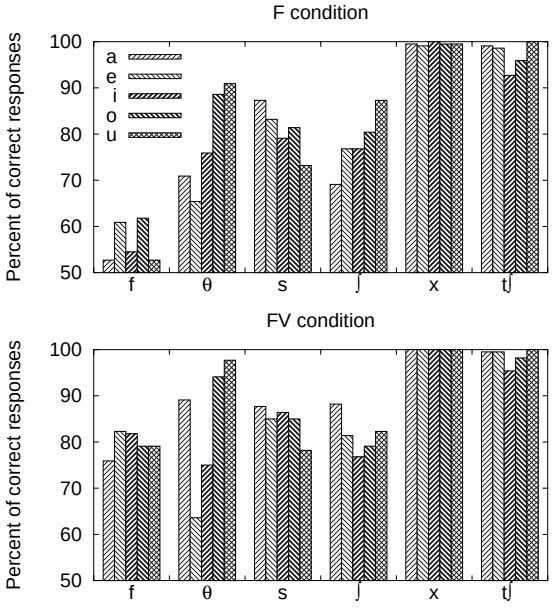

Figure 1: Correct recognition scores (in %) separately by vowel context. Top: F condition; bottom: FV condition.

was conducted for the responses of the FV condition. Again, the results show a significant main effect of fricative ($F(5, 594)=23.574$, $p<0.005$), plus a significant fricatives × vowel interaction ($F(20, 579)=2.558$, $p<0.0005$). The data were again submitted to a one-way analysis of variance and the Scheffé test (significance $p<0.05$) with the fricative as the independent factor ($F(5, 594)=22.270$, $p<0.00005$). The Scheffé test now produced a different grouping of fricatives: Subgroup 1 (/θ, f, s, ʃ/); Subgroup 2 (/tʃ, x/). In the FV condition /tʃ/ and /x/ are better identified than the rest, which have similar identification rates. Next, each phoneme was considered separately, and the data was submitted to a one-way analysis of variance and the Scheffé test (significance $p<0.05$) with the vowel as the independent factor. Again the effect of the vowel was only significant for the fricative /f/ ($F(4, 95)=12.568$, $p<0.0005$) with the following subgroups: Subgroup 1 (/e, i/); Subgroup 2 (/i, a/); Subgroup 3 (/a, o, u/). Again, /f/ is best recognized in the context of /u/ and /o/, and worst recognized in the context of /e/ (see figure 1). A one-way analysis of variance for each frica-

e-rthabla, N1, Abril 2003, ISSN: 1695-9914

	/f/	/θ/	/s/	/ʃ/	/x/	/tʃ/	another
/f/	78.4	19.8	0.2	0.0	1.5	0.0	0.1
/θ/	38.2	56.5	2.4	0.3	1.6	0.7	0.3
/s/	0.4	1.7	80.8	15.6	0.0	1.4	0.0
/ʃ/	0.2	0.4	16.8	78.4	0.0	4.0	0.2
/x/	0.4	0.1	0.0	0.0	99.5	0.0	0.0
/tʃ/	0.0	0.3	0.7	1.7	0.0	97.3	0.0
/f/	83.9	14.7	0.0	0.0	1.4	0.0	0.0
/θ/	17.0	79.6	0.4	0.2	1.4	0.7	0.6
/s/	0.5	3.0	84.4	10.4	0.1	1.4	0.1
/ʃ/	0.1	1.1	9.4	81.5	0.0	7.7	1.1
/x/	0.0	0.0	0.0	0.0	100.0	0.0	0.0
/tʃ/	0.0	0.2	0.2	1.1	0.0	98.5	0.0

Table 1: Confusion matrices (in %) obtained from the listeners' responses. Top: F condition (isolated fricative noise); bottom: FV condition (fricative noise plus 51.2 ms of the following vowel).

tive with the sex of the speaker as the independent factor showed a significant effect only for fricative /s/ ($F(1, 98)=4.130$, $p<0.045$).

The perceptual data of both experiments was submitted to a one-way analysis of variance in order to determine if the different conditions (F vs. FV) affect the listeners' responses. The effect of the experiment was significant ($F(1, 1198)=23.104$, $p<0.0005$). Then, each fricative was analysed separately searching for the same effect. Only for fricative /θ/ the difference between the two conditions was significant ($F(1, 198)=57.366$, $p<0.005$): listeners identified more consistently the fricative /θ/ in the presence of the following vowel than in the isolated condition (79.6% vs. 56.5%).

It is evident from the confusion matrices of both experiments that most confusions occur between the pairs /f-θ/ and /s-ʃ/. It has been reported in the literature that the identification of both pairs is affected by the vocalic portion associated with the consonant (Harris, 1958; La Riviere et al., 1975). This question was investigated more deeply.

The effect of the experiment was investigated in each of the ten fricative × vowel combinations formed by each of the fricatives /θ/ and /f/ and each of the five vowels. A one-way analysis of variance revealed a significant effect of the experiment for /f/ in the context of /a/ ($F(1, 38)=11.6564$, $p<0.0015$), /f/ in the context of /u/ ($F(1, 38)=4.4764$, $p<0.0410$) and for fricative /θ/ in all contexts. Then, /θ/ is equally well identified in all vocalic contexts, and identification in the FV condition is significantly higher than in the F condition. Identification of /f/ is affected by the vocalic context, improving in the FV condition with respect to the F condition only in the context of /a/ and /u/.

According to several studies (Whalen, 1981; Johnson, 1991; Yeni-Komshian and Soli, 1981; Mann and Repp, 1980), both the vocalic context and the sex of the speaker affect the acoustic characteristics of the fricatives /s/ and /ʃ/ and, as a consequence, may affect their auditory identification as well. A two-way analysis of variance (vowel × sex) was performed for /s/ and /ʃ/ separately in both conditions (F and FV). There was only a significant main effect of sex for /s/ in the FV condition ($F(1, 98)=4.13$, $p<0.045$). Since there was an effect of the sex of the speaker for /s/, the effect of the experiment was investigated in each of the twenty fricative × vowel × sex combinations, formed by each of the fricatives /s/ and /ʃ/ and each of the five vowels, for men and women separately. The one-way analysis of variance revealed a significant effect of the experiment only for /s/ in the context of /i/ pronounced by women. In view of these results, neither the vocalic context nor the sex of the speaker cause a significant effect on the recognition of either /s/ of /ʃ/, and there is no improvement in the FV condition with respect to the F condition except for the women's /s/ in the presence of /i/ (89.1% in the F condition vs. 97.3% in

e-rthabla, N1, Abril 2003, ISSN: 1695-9914

the FV condition).

The results of the perceptual experiment indicate that the auditory identification of fricatives is affected by vocalic context in few cases, particularly the case of /f/ in both the F and FV conditions. Contrary to the results of Yeni-Komshian and Soli (1981), who found that isolated fricative noises are most accurately identified when they are excised from the /a/ context, the best results for /f/ are obtained in the context of /o/ and /u/ in both the F condition (88.6% and 90.9%, respectively) and in the FV condition (94.1% and 97.7%, respectively), while the lower identification rates are obtained by /e/ in both the F condition (65.4%) and the FV condition (63.6%). For the rest of the fricatives there was no evidence of differences in fricative identification due to the vocalic context in either condition. It should be mentioned that in the study of Yeni-Komshian and Soli (1981) only the fricatives /s, ʃ, z, ʒ/ were studied.

In the /s-ʃ/ distinction, no vocalic context effects showed up. This is in contrast with other studies (Whalen, 1981; Johnson, 1991; Mann and Repp, 1980) claiming that the perceptual identification of the fricative noises of /s/ and /ʃ/ in isolation is affected by the vocalic context and the sex of the speaker.

The only clear effect of adding the vocalic context to the fricative noise arose for fricative /θ/ in all vocalic contexts and for fricative /f/ in the context of /a/ and /u/. Pittman and Stelmachowicz (2000) studied the importance given to vowel, transition and fricative segments, by a group of normal and hearing-impaired listeners, finding that the transition regions of /θ/ were heavily weighted. For /f/, the three segments were equally weighted, whereas for /s/ and /ʃ/ the fricative segments were heavily favoured by the listeners. The results of our experiments confirm the perceptual role played by the vowel in the /f-θ/ distinction.

Finally, /x/ and /tʃ/ were the phonemes with the highest recognition rates. Place of articulation of /x/ (velar) is fairly different from the places of articulation of the other phonemes, and as a result its distinct spectral shape makes auditory identification straightforward. The spectra of the palato-alveolar fricative and affricate, /ʃ/ and /tʃ/, show similar characteristics only after the affricate release (Stevens, 1993). Thus, manner, encoded as the spectral changes occurring during the first milliseconds including affricate release, probably provided an additional cue that helped listeners in identifying the voiceless affricate.

3 Acoustic analysis

The acoustic analysis was performed on exactly the same stimuli that were used in the perceptual experiments, i.e. on the whole fricative noise for the F condition and on the fricative noise plus 51.2 ms of the following vowel for the FV condition. The relative amplitudes of the friction noise and the vowel were kept exactly as in the perceptual experiments.

Feijóo et al. (1999) found significant differences in the classification of fricatives using spectra extracted from the initial, central and final part of the noise. Therefore a dynamical spectral representation was used in the present study. The spectral features were computed for several speech frames samples every 5 ms. The number of frames depended on the duration of each particular fricative, so it varied from token to token. Windows of different lengths (10, 15 and 20 ms) were investigated.

3.1 Acoustic parameters

A previous study (Feijóo et al., 1999) comparing the performance of several groups of parameters revealed that the acoustic characteristics of the fricatives can be represented by a set of FFT-derived cepstral coefficients. Linear cepstrum obtained better results than mel cepstrum or spectral moments in the classification of the tokens. The method used to obtain the linear cepstrum is the same of Davis and Mermelstein (1980). The speech signal is windowed using a Hamming window and a 1024 points FFT is computed for each frame. If Y_k represents the k-th coefficient of the log magnitude spectrum, then the i-th cepstral coefficient is computed as:

$$LFCC_i = \sum_{k=0}^{N-1} Y_k \cos\left(\frac{\pi i k}{N}\right) \quad i = 1, 2, \ldots, M \quad (1)$$

where M is the number of cepstral coefficients and N is the number of FFT magnitude coefficients.

The cepstral coefficients were computed over the whole frequency range of the sampled signals, i.e. between 100 and 9200 Hz.

Since the cepstral parameters are computed over several frames of speech, in order to encode the

e-rthabla, N1, Abril 2003, ISSN: 1695-9914

dynamic spectra their values where combined using the method proposed by Nossair and Zahorian (1991) for the classification of stop consonants. Preliminary experiments showed that three coefficients gave the best results in combination with the linear cepstrum. The value of each cepstral coefficient is expanded over all the frames analysed using a cosine basis-vector expansion:

$$LFCC_i(n) = \sum_{k=1}^{M} C_k \cos \frac{(k-1)\pi(n-0.5)}{L} \quad (2)$$

where $LFCC_i(n)$ is the i-th cepstral coefficient of frame n, M is the number of cosine coefficients used for expansion, C_k is the k-th cosine coefficient and L is the total number of frames. Therefore, the total number of parameters used to encode the dynamic spectra is equal to the number of cepstral coefficients $\times$ 3 cosine coefficients.

3.2 Classification and APP scores

The method used for classifying the acoustic tokens was a classical *Quadratic discriminant function* based on the Bayes' rule. For an individual token, *a posteriori probabilities* of membership in each group are calculated from the distance between the token and the centroids using Bayes' theorem:

$$P(w_i|X) = \frac{P(X|w_i)P(w_i)}{\sum_{i=1}^{n} P(X|w_i)P(w_i)} \quad (3)$$

where X is the vector representative of the token to be classified; w_i is the vector representing the i-th classification group; n is the number of possible groups; $P(w_i)$ is the *a priori probability* for each group; $P(X|w_i)$ are the *conditional probability functions*, obtained assuming that the token in question belongs to one of the groups, and computing the probability that the value of the quadratic discriminant equation predicts that the token belongs to that group; and $P(w_i|X)$ is the *a posteriori probability* that X belongs to the i-th group characterized by w_i.

Each token is classified as a member of the group for which the *a posteriori probability* is higher.

Classification of the samples is carried out using the leave-one-out method, for which the samples to be classified are different from the samples used to train the classifier, obtaining a closer estimate of the true classification rate (Fukunaga, 1972).

3.3 Relation between acoustic measurements and fricative identification by listeners

The use of the *a posteriori probabilities* as acoustic distances is based on the fact that the *a posteriori probability* of a token in a certain group should be the highest among the possible groups when the token actually belongs to that group. Each token, then, is represented by as many APP scores as groups. In the present case, the i-th token is associated with 6 probabilities, $\mathrm{APP}^i(r)$, where $r = 1, \ldots, 6$, represents each reference group (*distance profile*), and $i = 1, \ldots, 600$. The use of APP is based on the fact that auditory identification is influenced by the size and nature of the set of possible classes (Neary and Hogan, 1986).

In the same way that each token is associated with six acoustic-phonetic distances, it must be also associated with six auditory distances, one for each perceptual reference class. For each token, a *response profile* is computed as the proportion of responses assigned to each of the six response classes by the listeners. The response probabilities are estimated by pooling responses over listeners, under the basic assumption that the differences in listening abilities across listeners are random (Assman and Summerfield, 1989). The *response profile*, then, consists of six auditory distances, $\mathrm{AD}^i(r)$, $r = 1, \ldots, 6$, $i = 1, \ldots, 600$.

Auditory and acoustic distances were compared using the product-moment Pearson correlation for both the F and FV conditions. Unless indicated, all correlations are significant ($p < 0.001$). Prior to calculation of the correlation, both groups of data were transformed to have zero means (Bendat and Piersol, 1971).

Two different correlations were considered. The first one (which shall be denoted as *overall correlation*) is similar to that described in Assman and Summerfield (1989): the *response profiles* are correlated with the *distance profiles*, and one correlation coefficient is obtained over all the possible responses. The second one consists in the computation of one correlation coefficient in each of the 6 classes (denoted as *class correlation*).

e-rthabla, N1, Abril 2003, ISSN: 1695-9914

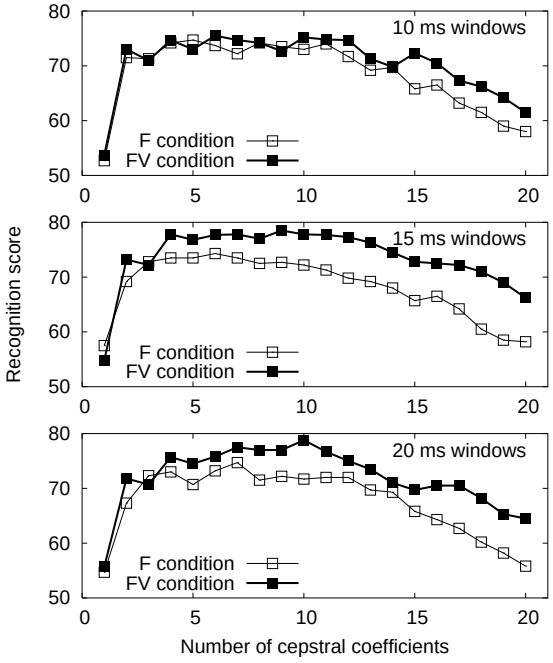

Figure 2: Correct recognition scores (in %) obtained by the cepstral coefficients for the F and FV conditions as a function of the number of coefficients. Top: 10 ms window; middle: 15 ms window; bottom: 20 ms window.

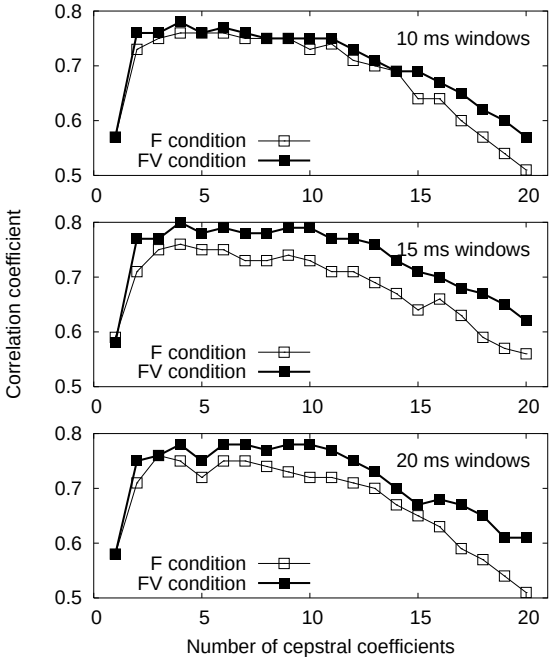

Figure 3: Results of the *overall correlation* between the APP and the *response profiles* for the F and FV conditions, as a function of the number of coefficients. Top: 10 ms window; middle: 15 ms window; bottom: 20 ms window.

3.4 Results and discussion

Figure 2 shows the classification results for the F and FV conditions using windows of 10, 15 and 20 ms. The best results are obtained using either 15 or 20 ms long windows. In both cases there is a clear improvement in the FV condition with respect to the F condition, indicating that the method is able to extract spectral characteristics relevant for fricative distinction from the vocalic part. The best results in the F condition are obtained by the first 5 cepstral coefficients for the 10 ms window (74.7% correct); by the first 6 coefficients for the 15 ms window (74.3% correct); and by the first 7 coefficients for the 20 ms window (74.7% correct). In the FV condition the best results correspond to the first 6 coefficients for the 10 ms window (75.5% correct); to the first 9 coefficients for the 15 ms window (78.5% correct); and to the first 10 coefficients for the 20 ms window (78.8% correct).

The results of the *overall correlation* between the observed and predicted responses as a function of the number of coefficients for the different windows are shown in figure 3. In the F condition the best results are obtained with the first 4–6 cepstral coefficients for the 10 ms window (0.76, $p < 0.001$); with the first 4 coefficients for the 15 ms window (0.76, $p < 0.001$); and with the first 4 coefficients for the 20 ms window (0.76, $p < 0.001$). In the FV condition, the best correlations correspond to the first 4 coefficients for the 10 ms window (0.78, $p < 0.001$); to the first 4 coefficients for the 15 ms window (0.80, $p < 0.001$); and to either the first 4, 6, 7, 9 or 10 coefficients for the 20 ms window (0.78, $p < 0.001$).

Overall, the results show that the acoustic characteristics of fricatives in the FV condition are better defined than in the F condition. Both the correct classification rates and the correlation between the observed and predicted responses in the FV condition are higher than in the F condition. Both results support the importance of the role played

e-rthabla, N1, Abril 2003, ISSN: 1695-9914

	/f/	/θ/	/s/	/ʃ/	/x/	/tʃ/
/f/	63	24	1	1	5	6
/θ/	28	52	7	2	3	8
/s/	0	4	68	26	0	2
/ʃ/	0	1	15	80	2	2
/x/	3	7	0	0	90	0
/tʃ/	2	4	4	2	0	88
/f/	61	35	1	0	2	1
/θ/	26	61	3	0	2	8
/s/	0	3	75	21	0	1
/ʃ/	0	0	17	83	0	0
/x/	3	4	0	0	93	0
/tʃ/	0	3	2	1	0	94

Table 2: Confusion matrices (in %) obtained by the first 4 cepstral coefficients (15 ms window). Top: F condition; bottom: FV condition.

by the vocalic part in fricative identification, even for fricatives for which the noise is a strong cue. The method employed is able to extract from the vocalic part information relevant for fricative distinction. Nevertheless, the automatic classification scores in both conditions are lower than those obtained by the listeners (81.8% in the F condition, and 88.0% in the FV condition). Moreover, the highest correlation scores do not correspond to the variables that obtain the best classification rates. If the degree of correlation is considered as the true indicator of the acoustic characteristic most likely used by the listeners in the auditory identification, then the first 4 cepstral coefficients computed from the 15 ms windows would represent the best spectral description, despite the fact that their classification rates are not the best ones.

Table 2 shows the confusion matrices obtained by the first 4 cepstral coefficients computed from the 15 ms window (73.5% correct in the F condition, and 77.8% correct in the FV condition). Only a marginal improvement is obtained for /θ/ in the FV condition (52% in the F condition; 61% in the FV condition). There is a slight decrease in classification performance for /f/ in the FV condition (63% in the F condition; 61% in the FV condition). For both /θ/ and /f/ the classification scores obtained by the variables in the FV condition are well below those of the listeners. That indicates that the automatic method is unable to extract from the vocalic portion of /f/ and /θ/ enough information about the place of articulation of both fricatives.

Some coincidences also show up. The confusion pattern of the listeners and of the acoustic method are similar: confusions generally take place between /f-θ/, and between /s-ʃ/ in both conditions. The rough spectral characterization provided by the 4 cepstral coefficients is accurate enough to reproduce some of the spectral and auditory similarities present in the signals. The dynamic spectra are also able to provide enough manner cues for the /tʃ/ distinction, especially in the FV condition, despite the spectral similarities reported between /tʃ/ and /ʃ/ (Stevens, 1993).

Results of the *class correlation* corresponding to the first 4 cepstral coefficients computed from the 15 ms window are shown in figure 4 (top), for the F and FV conditions. There is a general improvement in the correlation for the FV condition, as expected. The best correlations in both conditions are obtained by /x/ and /tʃ/, which is not surprising, since both the listeners and the acoustic method obtained high identification rates for both phonemes. For /s/ and /ʃ/ the classification rates of the listeners and the acoustic variables in the FV condition are similar, but the correlation coefficients of /s/ (0.80) and /ʃ/ (0.83) are much lower than those of /x/ (0.94) and /tʃ/ (0.92). This means that a similarity in classification rates is not a very strong indication of the relationship between the acoustic classification and the perceptual evaluation. The present results show that the dynamic spectra encoded by the cosine coefficients are able to capture only until some extent, the acoustic characteristics that bear the relevant auditory information.

Lastly, it might happen that the acoustic method is not sensitive enough to reflect the fine differences in the values of the pooled listeners' responses. To investigate this aspect, a *winner-takes-all* score was computed. The *winner-takes-all* (WTA) represents the proportion of stimuli for which the acoustic method's most likely response is equal to the actually most-occurring response (Smits et al., 1996). Figure 4 (bottom) shows the outcome of the analysis separately for each phoneme. The WTA levels for the different fricatives/affricate show the same trend observed for the correlation coefficients.

Very distinct phonetic categories that are easily recognized by the listeners may be represented by a low dimensional acoustic characterization. For

e-rthabla, N1, Abril 2003, ISSN: 1695-9914

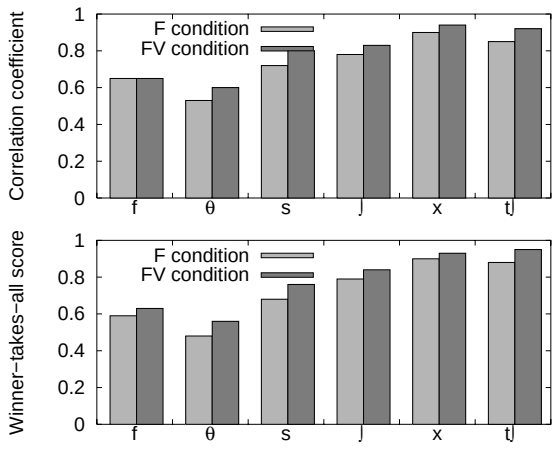

Figure 4: Comparison between perception and acoustic model. Top: *Class correlation* results between the APP scores and the *response profiles*; bottom: Winner-takes-all scores. Results are presented for each phoneme in the F and FV conditions. The acoustic analysis corresponds to the first 4 cepstral coefficients obtained with the 15 ms window.

instance, separation of /s/ and /θ/ may be accomplished in terms of a couple of single variables like Energy and Zero-crossings (Feijóo et al., 1994). That does not mean that listeners do not process many other acoustic cues that are present in the signal. It is dubious that those same variables could, for instance, separate /θ/, /f/ and /s/. The utility of the other cues may show up when the circumstances in which the recognition takes place are more difficult. The importance of the additional acoustic cues does not have to be constant across the phonetic categories. The global spectral characterization provided by the 4 cepstral coefficients is clearly not sufficient to describe equally well all the phonemes considered in our study, although it still provides a reasonable characterization of those sounds.

The dynamic properties that encode the fricative/affricate distinction were acceptably well represented by the method. Despite the reported similarities between /ʃ/ and /tʃ/, it is obvious that listeners were able to identify the affricate without problem, using manner as an additional cue. The acoustic method was efficient enough to represent the manner cue, probably because the dynamic spectra were accurate enough to encode the spectral changes that take place during the first milliseconds of affricate release.

The combination of spectra using the cosine coefficients could not completely reproduce the phonetic integration that took place when the vowel was added, particularly for /θ/ and /f/. The failure of the acoustic method in replicating the phonetic integration process for those phonemes may be attributed to a number of reasons: a) the acoustic variables that are auditorily integrated are not well represented by the global spectral shape; b) the order of the spectral analysis should be different for the noise and the vowel; c) the number of cosine coefficients used to encode the dynamic spectra across the FV segment is not adequate to represent the spectral changes that take place during the transition from an unvoiced sound (the fricative) into a voiced sound (the vowel); d) the statistical classifier is not adequate. Nossair and Zahorian (1991), though, obtained automatic recognition rates similar to those of the human listeners using three cosine coefficients for the classification of stop consonants, over segments of speech including the burst and the vowel. Further research is necessary to determine the importance of the above mentioned points.

4 Conclusions

The voiceless fricatives and affricates of Galician have been studied acoustically and auditorily in the F and FV conditions.

The results of the perceptual tests show that in the F condition, /θ/ is the fricative with the lowest recognition score, while /x/ and /tʃ/ are better identified than the rest. Adding the vowel caused a significant improvement in recognition for fricative /θ/ in all vocalic contexts, and for fricative /f/ in the contexts of /a/ and /u/. The effect of vocalic context on the perceptual identification was only significant for /f/ in both the F and FV conditions.

On the other hand, the results of the acoustic analysis show that the spectral characteristics of the voiceless fricatives and affricates are better defined in the FV condition than in the F condition, supporting the importance of the role played by the vowel in fricative identification. The best match

e-rthabla, N1, Abril 2003, ISSN: 1695-9914

between the auditory and acoustic representation was obtained by a low order spectral representation (4 cepstral coefficients). Nevertheless, the performance of the acoustic method was inferior to the listeners' performance in both conditions, especially for /f/ and /θ/. The results obtained for those fricatives indicate that the present method is unable to replicate the perceptual integration that takes place in the auditory identification of those phonemes. The acoustic model, though, is accurate enough to replicate the pattern of confusions of the listeners. The method is also able to provide a reasonable characterization of the cues for the /ʃ-tʃ/ distinction. For /θ/ and /f/ the global spectral description is not as accurate as for the rest of phonemes, probably because other variables are also important for that distinction.

The correlations between the *response profiles* and the APP scores proved to be a valid method for relating the perceptual responses to the acoustic representation, particularly the use of *class correlation*. The *class correlation* scores can be considered as a good indicator of the match between perceptual and acoustic spaces.

Acknowledgements

This research was supported by a grant from the University of Santiago given to Santiago Fernández. The authors would like to thank Susana Villaverde for her assistance with the perceptual experiments. This work was financed by Xunta de Galicia under project PGIDT00PXI20608PR.

References

P. F. Assman and Q. Summerfield. Modeling the perception of concurrent vowels: vowels with the same fundamental frequency. *J. Acoust. Soc. Am.*, 85:327–338, 1989.

S. Behrens and S. E. Blumstein. On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants. *J. Acoust. Soc. Am.*, 84:861–867, 1988.

J. S. Bendat and A. G. Piersol. *Random data: analysis and measurement procedures.* Wiley-Interscience, New York, 1971.

A. M. Borzone de Manrique and M. I. Massone. Acoustic analysis and perception of Spanish fricative consonants. *J. Acoust. Soc. Am.*, 69: 1145–1153, 1981.

S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech and Signal Proc.*, 28:357–366, 1980.

S. Feijóo and S. Fernández. Temporal effects of phonetic integration in CV syllables. In *Proc. Forum Acusticum*, Sevilla, 2002.

S. Feijóo, S. Fernández, N. Barros, and R. Balsa. Acoustic and perceptual characteristics of the Spanish fricatives. In *Proc. EuroSpeech*, Budapest, 1999.

S. Feijóo, J. Taboada, J. R. Fernández, and N. Barros. Automatic classification of the Spanish fricatives /θ/ and /s/. *Acustica*, 80:442–452, 1994.

P. Jr. Flipsen, L. Shriberg, G. Weismer, H. Karlsson, and J. McSweeney. Acoustic characteristics of /s/ in adolescents. *J. Speech, Lang. Hear. Res.*, 42:663–667, 1999.

K. Forrest, G. Weismer, P. Milenkovic, and R. N. Dougall. Statistical analysis of word-initial voiceless obstruents: preliminary data. *J. Acoust. Soc. Am.*, 84:115–123, 1988.

K. Fukunaga. *Statistical Pattern Recognition.* Academic Press, 1972.

K. S. Harris. Cues for the discrimination of American English fricatives in spoken syllables. *Lang. Speech*, 1:1–7, 1958.

W. Jassem. Classification of fricative spectra using statistical discriminant functions. In B. Lindblom and S. Ohman, editors, *Frontiers of Speech Communication*, pages 77–91. Academic Press, New York, 1979.

W. Jassem. An acoustical linear-predictive and statistical discriminant analysis of Polish fricatives and affricates. In W. Jassem, Cz. Babztura, and K. Jassem, editors, *Speech Language and Technology*, volume 2, pages 9–45. Polish Phonetic Association, Poznari, 1998.

K. Johnson. Differential effects of speaker and vowel variability on fricative perception. *Lang. Speech*, 34:265–279, 1991.

A. Jongman. Duration of frication noise required for identification of English fricatives. *J. Acoust. Soc. Am.*, 85:1718–1725, 1989.

A. Jongman, J. Sereno, R. Wayland, and S. Wong. Acoustic characteristics of English fricatives. *J. Acoust. Soc. Am.*, 108:1252–1263, 2000.

C. La Riviere, H. Winitz, and E. Herriman. The distribution of perceptual cues in English prevocalic fricatives. *J. Speech Hear. Res.*, 18:613–622, 1975.

V. A. Mann and B. H. Repp. Influence of vocalic context on perception of the /ʃ/-/s/ distinction. *Percept. Psychophys.*, 28:213–228, 1980.

T. M. Neary and J. T. Hogan. Phonological contrast in experimental linguistics: relating distributions of measurement of production data to categorization curves. In J. Ohala, editor, *Experimental phonology*, pages 141–161. Academic, London, 1986.

Z. B. Nossair and S. A. Zahorian. Dynamic spectral shape features as acoustic correlates for initial stop consonants. *J. Acoust. Soc. Am.*, 89:2978–2991, 1991.

L. C. Nygaard and D. B. Pisoni. Speech perception: new directions in research and theory. In J. L. Miller and P. D. Eimas, editors, *Speech, Language and Communication*, pages 63–95. Academic Press, San Diego, 1995.

J. P. Olive, A. Greenwood, and J. Coleman. *Acoustics of American English speech*. Springer-Verlag, 1993.

A. L. Pittman and P. G. Stelmachowicz. Perception of voiceless fricatives by normal-hearing and hearing-impared children and adults. *J. Speech, Lang. Hear. Res.*, 43:1389–1401, 2000.

L. R. Rabiner and R. W. Schafer. *Digital processing of speech signals*. Englewood Cliffs, New Jersey, 1978.

B. H. Repp. Integration and segregation in speech perception. *Lang. Speech*, 31:237–271, 1988.

C. H. Shadle and S. Mair. Quantifying spectral characteristics of fricatives. In *Proc. Int. Conf. Speech Lang. Proc.*, pages 1521–1524, Philadelphia, 1996.

R. Smits, L. ten Bosch, and R. Collier. Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. II. modeling and evaluation. *J. Acoust. Soc. Am.*, 100:3865–3881, 1996.

K. N. Stevens. Modeling affricate consonants. *Speech Comm.*, 13:33–43, 1993.

P. Strevens. Spectra of fricative noise in human speech. *Lang. Speech*, 3:32–49, 1960.

D. H. Whalen. Effects of vocalic formant transitions and vowel quality on the English /s/-/ʃ/ boundary. *J. Acoust. Soc. Am.*, 69:275–282, 1981.

D. H. Whalen. Perception of English /s/-/ʃ/ distinction relies on fricative noises and transitions, not on brief spectral slices. *J. Acoust. Soc. Am.*, 90:1776–1785, 1991.

G. H. Yeni-Komshian and S. D. Soli. Recognition of vowels from information in fricatives: perceptual evidence of fricative-vowel coarticulation. *J. Acoust. Soc. Am.*, 70:966–975, 1981.

F. Zeng and C. W. Turner. Recognition of voiceless fricatives by normal and hearing impaired subjects. *J. Speech Hear. Res.*, 33:440–449, 1990.