LOCATION OF SPECTRAL CUES FOR THE PERCEPTUAL IDENTIFICATION OF FRICATIVES

PACS REFERENCE: 43.71.Es

Feijóo, Sergio; Fernández, Santiago
University of Santiago de Compostela
Departamento de Física Aplicada, Facultad de Física
15782 Santiago de Compostela
SPAIN
Phone:+34 981563100 ext.14044
Fax: +34 981520676
E-mail: fasergio@usc.es

ABSTRACT
The perceptual space (MDS) obtained from the identification of low-pass filtered Spanish fricatives at different cutoff frequencies, has been correlated with the acoustic space computed from the energy level in spectral bands extracted with triangular mel filters, logarithmic rectangular filters and linear rectangular filters. The weights attached to every spectral band by the acoustic analysis were studied in order to determine which spectral bands contribute to the identification of fricatives and what is the minimum bandwidth required for a satisfactory characterization of fricatives. Some differences among the three acoustic representations showed up. Results show that the most important cues for the identification of fricatives are located below 6 kHz, and that the relation among frequency bands around 3 kHz and above 6 kHz provide finer detail for the /s-sh/ and /th-f/ distinctions.

INTRODUCTION

The spectral properties of the fricative noises seem to be the primary acoustic cue to the perceptual identification of isolated fricatives [1]. Some other acoustic characteristics of the fricative noise, such as duration and amplitude, have been studied. Those properties do not play a major role in the identification of place of articulation of fricatives, although amplitude may serve as a cue to the sibilant versus non-sibilant distinction in some cases. The spectral characteristics of Galician fricatives are similar to those reported for English fricatives. In general, it is considered that /s/ has spectral energy around 4 and 5 kHz (and above those frequencies), /sh/ around 2.5 and 3 kHz, while /f/ and /th/ have flat spectra with some high frequency peak (around 8 kHz for /f/). Some differences related to the sex of the speaker and to the quality of the neighboring vowel have also been reported, particularly for /s/ and /sh/.

The location of the most prominent spectral peak of fricatives has been used for classification tasks despite of being speaker and vowel dependent [2]. It was found that /s/ and /sh/ can be accurately classified on the basis of this spectral property. Non-sibilant fricatives, on the other hand, have not clearly dominating peaks and thus it was not possible to discriminate accurately /f/ and /th/. In this sense, a gross characterization of the spectra of fricatives is interesting since the spectral prominences observed in their spectrum are much broader than those of vowels and spectral energy spreads over a wide range of frequencies. Besides, the high variability observed in the detailed characteristics of the spectrum of the fricatives noise makes the use of detailed cues questionable.

This paper deals with the location of spectral cues for the perceptual identification of fricatives. Stimuli are low-pass filtered fricative noises. Cutoff frequencies were 16, 11, 8, 5.5, 4 and 3 kHz. The perceptual space obtained from a Multidimensional Scaling Analysis (MDS) on listeners' responses was compared with the acoustic space obtained from a Principal Components Analysis (PCA) on energy level in spectral bands. The results obtained with three different filterbanks are compared since each filterbank extracts different spectral properties. We focus our attention on the analysis of the weights attached to the energy levels by the PCA on each of the three filterbank outputs in order to clarify which spectral cues to place of articulation are relevant for fricative's identification. The use of stimuli with several bandwidths allows us to see how the absence of some of those cues influences the perceptual results.

MATERIALS AND METHOD

Tokens are the fricative noises of a set of FV syllables formed by the Galician voiceless fricatives /f, th, s, sh/ with /a, e, i, o, u/. Two native speakers of Galician (Northwest Spain), one male and one female, pronounced the syllables carefully but in a natural way. While the use of a large data set would have provided us with a broad view of the processes involved, it nevertheless obscures particular details that may be important. A reduced and selected set, on the other hand, allows us to access more detailed information. The signals were recorded with a sampling frequency of 32 kHz, filtered at the Nyquist frequency, and normalized at 75% of the quantization range (16 bits). For the first condition the signals were high-pass filtered with a cutoff frequency of 100 Hz in order to eliminate undesired low-frequency noises. For the following conditions they were band-pass filtered with a lower cutoff frequency of 100 Hz and upper cutoff frequencies of 11, 8, 5.5, 4 and 3 kHz. The attenuation in the stopbands was approximately 70 dB. Fricative noises were extracted from the syllables and the beginning and end of the segments were smoothed with a 10 ms long cosine-type window to prevent undesired clicks. The stimuli are the 240 fricative noises=4 fricatives x 5 vowels x 2 sexes x 6 bandwidths.

37 Galician speaking students carried out the perceptual experiments for course credits. Stimuli were presented at random to limit the loudness cues that might be present if each low-pass filter condition was tested separately. Subjects were provided with five different options, /f, th, s, sh/ and *another sound*, in order to avoid guessing. A MDS (asymmetric individual differences) analysis on the six confusion matrices (one for each cutoff frequency) was carried out. For the acoustic analysis, the energy level in spectral bands covering the whole spectrum (up to 16 kHz) of the fricative noises was computed. The inclusion of filtered bands in the acoustic characterization allow us to assess whether listeners have used information from the filtered region or not. Three filterbanks were used: 1) a mel filterbank (MFB), 2) the logaritmic output of a set of rectangular, 1 kHz –wide, non overlapped and linearly spaced spectral bands (log-LFB), and 3) the linear output of the second filterbank (LFB). The MFB is assumed to agree best with the auditory process. The log-LFB has a resolution in frequency which is supposed to agree better than the MFB with the acoustic characteristics of fricatives. The LFB allows us to evaluate to what extent the perceptual results can be explained on the basis of only the most important concentrations of spectral energy. A PCA was carried out on the acoustic data for each cutoff frequency and acoustic characterization. PCA selects those linear combinations of the original variables explaining a larger percentage of the variance.

RESULTS

Analysis of the perceptual results was presented in a previous study [3]. Basically, results show that the fricatives /f/ and /sh/ are the least influenced by the bandwidth reduction. Confusions take place between /f/ and /th/, and between /s/ and /sh/, although for the lowest cutoff frequecies /s/ is also confused with both /th/ and /f/. No significant differences between speakers/sexes showed up. The MDS analysis represents the stimuli spatially, taking into account that the perceptual distances reflected in the confusion matrices must be kept. A two-

dimensional space gave interpretable results, explaining 99.5% of the variance. PCA was carried out for each condition separately, since overall spectral differences among diferent cutoff frequencies are not of interest. The first two principal components explained most of the variance in the sample: around 53% for the MFB, 85% for the log-LFB and 94% for the LFB. The percentage of explained variance increases as the bandwidth decreases, except for the MFB, for which the opposite is true. Both log-LFB and LFB seem to be more compact acoustic characterizations of the fricative noise than MFB. MFB seems to have a somewhat inadequate frequency resolution of 100 Hz at frequencies below 1 kHz, where little or non important information for fricatives has been reported (see introduction). The perceptual and acoustic spaces are shown in figure 1. The centroids of the distributions for each fricative and cutoff frequency have been plotted. These distributions consist of 10 tokens= 2 speakers x 5 vocalic contexts. Canonical correlations (r^2) between the perceptual and the acoustic spaces are: 0.70 for the MFB, 0.87 for the log-LFB and 0.75 for the LFB. The location of the different phonetic categories is well modelled by the acoustic analysis. Fricatives in the acoustic space are particularly well separated when MFB outputs are used (in particular, notice the /f/-/th/ distinction), although there is not a particularly good agreement with the perceptual space, since /f/ and /th/ are perceptually ambiguous. LFB outputs does not show almost any variation with the spectral bandwidth. Log-LFB outputs are more in agreement with the perceptual space than the other two acoustic characterizations.

Let us take a look at the weighting of the energy level in spectral bands performed by the analysis in order to gain an insight into the spectral cues relevant to the identification of the place of articulation of fricatives. Mathematically, the weighting is done by the PCA according to the following equation:

$$\mathrm{PC}_j = \sum_i \mathrm{weight}_i \left(\mathrm{band}_i - \mathrm{mean}_i\right)$$

where band_i is the energy level of the *i*th spectral band, mean_i is the mean energy level of band *i* for all tokens entered into the PCA, weight_i is the weight assigned by the analysis to the *i*th spectral band and PC_j is the *j*th principal component. The weights are shown in figure 2 for each

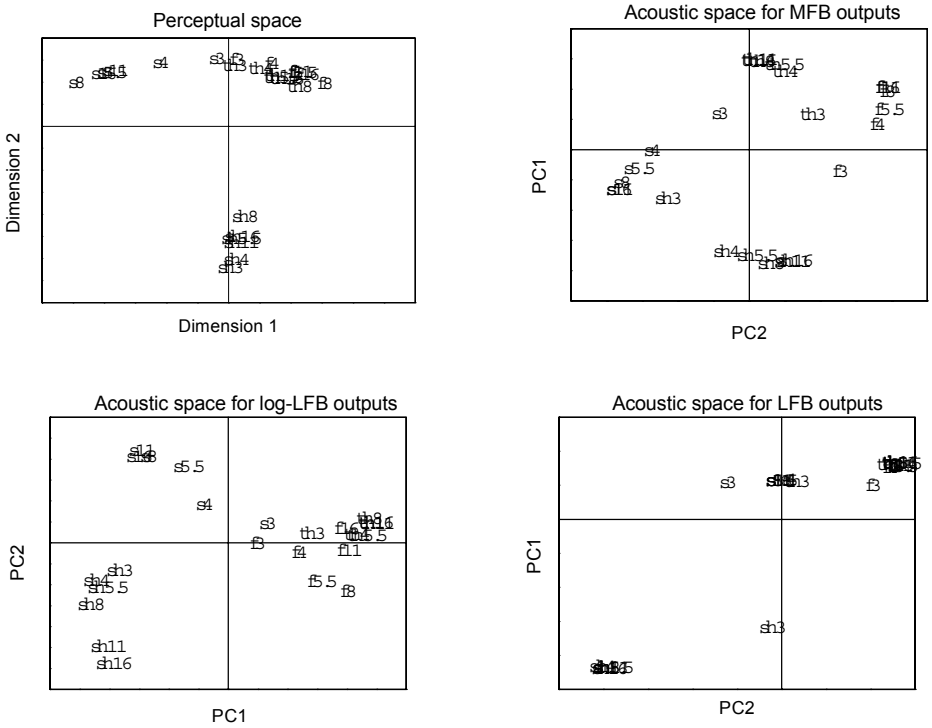

Figure 1. Perceptual space (top, on the left) and acoustic spaces for the three acoustic characterizations. The location of each fricative for each cutoff frequency is represented by the name of the fricative (/f/, /th/, /s/ or /sh/) plus the cutoff frequency (16, 11, 8, 5.5, 4 or 3 kHz). The horizontal and vertical lines mark the origin of each coordinate.

filterbank and condition separately. To interpret these results the spectra of fricatives for every filterbank were inspected.

For the MFB and a bandwidth of 16 kHz, the second principal component (PC2) distinguishes every pair of fricatives except /th/ from /sh/ (see figure 1). /f/ is given a positive value, therefore, as the weighting of the energy levels indicate (see figure 2), /f/ is characterized by a marked concentration of energy at low frequencies (below 3 kHz). On the other hand, /s/ is negative due to the presence of spectral energy around 6 kHz. For /th/ and /sh/, both concentrations of spectral energy are compensated. The spectrum of /th/ shows some amount of spectral energy below 3 kHz but it also shows spectral energy around 6 kHz. Therefore, /th/ is slightly negative. The spectrum of /sh/ shows a great concentration of spectral energy around 3.5 kHz. The tails of that distribution compensate each other, /sh/ being slightly positive. The first principal component (PC1) distinguishes primarily sibilants from non-sibilants by means of the characteristic concentration of spectral energy that sibilants have around 4 kHz (especially /sh/) and above. For both principal components, weights above 8 kHz are almost zero. Therefore, as the bandwidth is decreased weights do not vary much. Most of the spectral properties used to distinguish among fricatives are still available for a cutoff frequency of 5.5 kHz, and even for a cutoff frequency of 4 kHz. For 3 kHz, the acoustic space is quite different, but it is still possible to discriminate among the four fricatives on the basis of the two principal components. The sibilant versus non-sibilant distinction (PC2) is carried out taking into account the relation between the concentration of spectral energy at low frequencies (below 2 kHz), which is large for non-sibilants, and the concentration of spectral energy above 2 kHz, which is large for sibilants. The distinction between /th, s/ and /f, sh/ (PC1) is due to the higher concentration of spectral energy below 3 kHz for the /f, sh/ pair. Summarizing, the MFB allows to discriminate among the four fricatives; nevertheless, this is not in accordance with listeners' perception, particularly in the case of the /th-f/ pair, which is perceptually ambiguous, and in the case of the lower cutoff frequency conditions which are also perceptually ambiguous.

The weights assigned by the PCA to the log-LFB outputs are shown in figure 2. The corresponding acoustic space attained the highest correlation with the perceptual space (r^2=0.87). For a cutoff frequency of 16 kHz, PC1 distinguishes between sibilants and non-sibilants (see figure 1). The weights (see figure 2) show us that if a particular fricative has a large concentration of spectral energy around 4 kHz which disminishes continuously as the frequency increases (this is the case for /s/ and /sh/), it will have a negative value for this particular component indicating that it is a sibilant fricative. If the main concentration of energy of a fricative is around 1 kHz (such as /f/ and /th/), it will be positive and, therefore, the fricative is non-sibilant. PC2 distinguishes primarily betwen /sh/ and the other fricatives. The weights evaluate the spectral content around 3 kHz with respect to the spectral content at high frequencies (above 8 kHz), negative weight being assigned to the former and positive weight being assigned to the last. If a fricative has a larger concentration of spectral energy around 3 kHz than above 8 kHz (/sh/), PC2 will be negative; otherwise it will be positive (/s/ and /th/). The highest positive values correspond to /s/, pointing to an important role of the high frequency content for this fricative and, to a lesser extent, for /th/. The opposite is true for /f/, for which the low energy content prevails over the high frequencies. As the bandwidth is reduced, spectral energy at high frequencies is no longer available, weights above the cutoff frequency becoming close to zero and weights below the cutoff frequency maintaining their values. For a cutoff frequency of 5.5 kHz the spectral cues used to distinguish among fricatives are still available. This coincides with listeners' perception. For lower cutoff frequencies both the acoustic and perceptual identity of the stimuli becomes ambiguous. The contribution of the spectral prominence around 4 kHz disminishes and /s/ is increasingly misidentified as /th/. Confusions with /f/ also increase because only low frequency energy is available now. On the other hand, /sh/ is still well identified because spectral energy around 4 kHz has been reduced in lesser extent than in the case of /s/. Note that for the narrower bandwidths, the contribution of high frequency spectral energy to PC2 becomes relevant around 5 kHz instead of 8 kHz. It seems that the analysis uses in those cases the available cues above 5 kHz to compensate for the lack of information above 8 kHz. Spectral energy above 6 kHz, approximately, seems to be important. Besides, for the narrower bandwidths the weights above the cutoff frequency for PC2 are not zero. An overall energy level of the filtered bands might help to distinguish place of

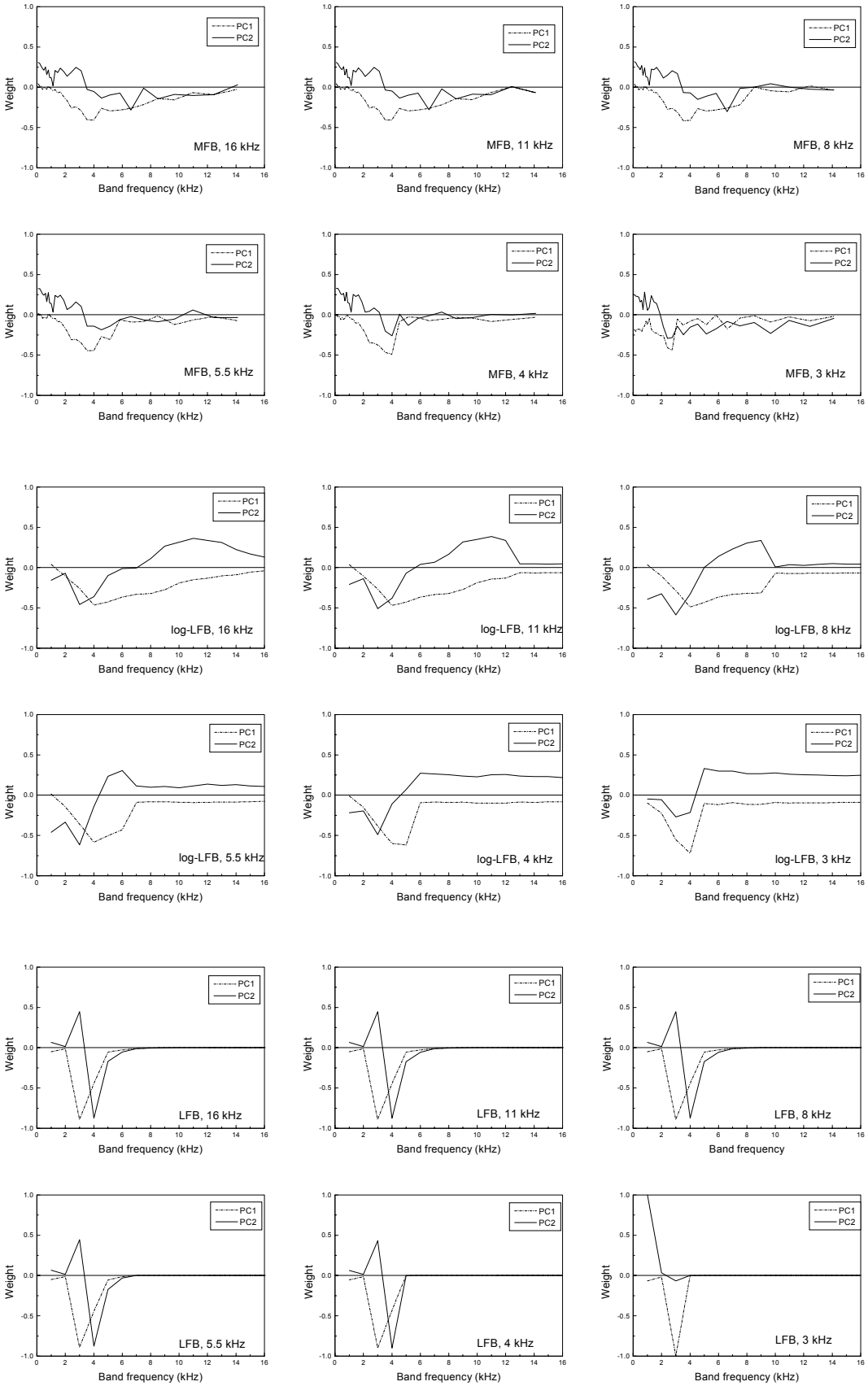

Figure 2. Weights attached by the principal components analysis to spectral bands for each of the three filterbanks: MFB (top), log-LFB (middle) and LFB (bottom); and cutoff frequency (16, 11, 8, 5.5, 4 and 3 kHz).

articulation in this case. It is possible that listeners use energy levels at frequencies higher than the cutoff frequency in the ambiguous conditions, possibly after performing a spectral integration of the energy included in the filtered bands, which means that the attenuation in the stop band was not sufficient.

For the LFB, weights are shown in figure 2. PC1 distinguishes among /sh/ and the other fricatives. This is achieved by means of giving negative weight to the spectral content at 3 kHz, weights for other frequency locations being zero. This coincides with the location in frequency of the main spectral prominence of /sh/. Since this information is available for every low-pass filtered stimuli, the weights are the same for every condition. PC2 distinguishes among non-sibilant fricatives, /s/ and /sh/. The weights take into account the relation between spectral energy around 4 kHz (negative weight) and spectral energy around 3 and 1 kHz (positive weight). For /sh/ spectral energy at 4 kHz is larger than spectral energy at 3 and 1 kHz. /sh/ has usually a well-defined trough at 3 kHz. This is not the case for /s/, for which spectral energy at 4 kHz compensates for spectral energy at 3 kHz. For non-sibilant fricatives, which have flat spectra, energy level at 1 kHz is higher than at other frequency locations. Therefore, PC2 is positive for non-sibilant fricatives. As the bandwidth is reduced, the acoustic space does not change, except for the lowest cutoff frequency (3 kHz). In this case, PC2 can not use information at 4 kHz; /s/, /sh/ and /th/ get closer and only /f/ is clearly distinguished because it has a larger concentration of energy around 1 kHz than the other fricatives. Summarizing, taking into account only the most prominent spectral concentrations and the relation among them, it is possible to reproduce the relation among fricatives in the perceptual space to some extent. Nevertheless, there is not enough information to reproduce the trajectory followed by fricatives in the perceptual space as the bandwidth is reduced, especially for the lowest cutoff frequencies. In this sense, PCA on log-LFB outputs obtains the best results.

CONCLUSIONS

For the perceptual identification of fricatives, the frequency region below 6 kHz is very important. As the bandwidth is reduced, /th/ and /s/ are the most affected fricatives while /f/ and /sh/ are affected to a lesser extent. For /f/ and /sh/, spectral energy up to 3 kHz contains most of the information required for their identification. These perceptual evidences can be explained from the energy level in spectral bands. MFB outputs did not agree satisfactorily with listeners' perception, nonetheless they discriminate among all four fricatives. LFB outputs corroborated that /s-sh/, but not /f-th/, can be discriminated on the basis of only the most prominent concentrations of spectral energy. Log-LFB outputs attained the best correlation (r^2=0.87) and offered a good explanation of the perceptual results: spectral energy below 6 kHz also seems to contain the most important cues for fricative identification. Sibilant fricatives are characterized by a concentration of spectral energy around 4 kHz, and non-sibilants by a concentration of energy around 1 kHz. The relation between the spectral energy around 3 kHz and above 6 kHz accounts for both the /s-sh/ and the /f-th/ distinctions.

ACKNOWLEDGMENTS

This work was financed by Xunta de Galicia under project PGIDT00PXI20608PR.

REFERENCES

1. LaRiviere, C., Winitz, H. and Herriman, E., "The distribution of perceptual cues in English prevocalic fricatives," J. Speech Hear. Res., 18:613-622, 1975.
2. Abdelatty Ali, A. M., van der Spiegel, J. and Mueller, P., "Acoustic phonetic features for the automatic classification of fricatives," J. Acoust. Soc. Am., 109:2217-2235, 2001.
3. Feijóo, S., Fernández, S. and Balsa, R., "Influence of frequency range in the perceptual recognition of fricatives," Proc. of Forum Acusticum (Berlin), Acustica-Acta Acustica, 85(1):S475(A).