CLASSIFICATION OF FRICATIVES

Marija Tabain and Catherine Watson
Speech, Hearing and Language Research Centre
Macquarie University

ABSTRACT - The purpose of this study is to explore voiceless fricative consonants in Australian English. In particular, attempts are made to classify the dental and the labio-dental fricatives ([T] and [f] respectively), using pre-emphasised averaged spectra of fricative tokens sampled at 44.1 kHz. Results suggest that the techniques used help to correctly classify the two non-sibilant fricatives, although the results are not as good as those for the other two fricatives, the alveolar [s] and the alveolopalatal [S].

INTRODUCTION

Whilst the acoustic characteristics of the sibilant fricatives of English, [s] and [S], are quite well understood, results from studies on the acoustic characteristics of the non-sibilant fricatives, [f] and [T], have proved less conclusive (Hughes & Halle 1956; Behrens & Blumstein 1988). The sibilant fricatives have a clearly identifiable peak below 10 kHz, whereas the non-sibilants present a more-or-less flat spectrum in this frequency range (Fant 1960; Flanagan 1972). Moreover, [f] and [T] have very low intensity, which combined with their lack of spectral peaks makes them very difficult to differentiate. Consideration of articulatory-to-acoustic mappings led us to believe that the smaller front cavity created between the upper teeth and the upper lip in the production of these fricatives might result in a higher frequency peak, above 10 kHz (Shadle 1985). The current study aims to explore the non-sibilant fricatives, looking in particular at information in the frequency range between 10 kHz and 20 kHz.

In order to look at the spectrum up to 20 kHz, we sampled our data at 44.1 kHz. We also chose to try pre-emphasis on our data in the belief that this would aid us in our investigation, since pre-emphasis has the effect of boosting the higher frequency components in the spectrum. As a further dimension, preliminary experiments suggested that the particular FFT technique used (averaged FFTs vs. spectral slice at the token midpoint) made a difference to the overall results. We therefore decided to explore this aspect of the spectral analysis in relation to fricatives as well.

As a result, the purpose of this work is to answer the following three questions:

i) is there a significant difference in the number of correctly classified fricative tokens depending on whether we use spectral slices at the midpoint or averaged FFTs in calculating our data?
ii) does pre-emphasis make a significant difference to the classification of [f] and [T]?
iii) does the higher sampling rate (i.e. 44.1 kHz) make a significant difference to the classification of [f] and [T]?

SPECTRAL CHARACTERISTICS OF [s], [S], [f] and [T].

Hughes and Halle (1956) examined fricative tokens (taken from isolated words) of both male and female speakers' productions of [s], [S] and [f]. They found that [f] had a relatively flat spectrum below 10 kHz, whereas [S] had spectral energy in the region 2 - 4 kHz, and [s] had spectral energy above 4 kHz. Strevens (1960) examined front (including [f] and [T]), mid (including [s] and [S]) and back fricatives, and found that the front fricatives were characterised by low intensity, smooth spectra; the mid fricatives by high intensity and significant peaks in the spectra; and the back fricatives by medium intensity and a marked formant-like structure. Gurlekian (1981) found

that the amplitude of the fricative, relevant to the following vowel, was important in the separation of [s] from [f], where [s] had greater and [f] lower amplitude relevant to the vowel. Behrens and Blumstein (1988) studied all four English voiceless fricatives, and (in contrast to Hughes and Halle for [f]) found a diffuse spread of energy between 1.5 and 8.5 kHz for both [f] and [T]. Shadle (1985) showed that the greater amplitude in [s] and [S] is due to the presence of an obstacle (namely, the lower teeth) some 3 cm downstream from the noise source (namely, at the constriction). This obstacle serves to increase the turbulence of the air-flow and to increase the amplitude. This is the characteristic feature of sibilant fricatives. The non-sibilant fricatives, by contrast, have no such obstacle, resulting in the very low energy levels which are their main characteristic.

METHOD

A database of fricative tokens taken from isolated word utterances was collected, comprising all eight types of English fricatives in CV position. Nine speakers were used (five male and four female). They were all speakers of General Australian English, and had no known speech or hearing difficulties. Recording sessions took place in a sound-treated studio under the direction of a technician. Each speaker read out a list of 717 different isolated words which were flashed up on a screen approximately 2 metres from the speaker. The word-list contained about 90 tokens of each fricative phoneme in syllable-initial (including word-initial) position. The words were both monosyllabic and polysyllabic, with varying stress patterns. Some words had variable stress patterns (e.g. conVICT vs. CONvict) -- in this case, both pronunciations were elicited. Speakers repeated the words if either they or the technician felt that this was necessary (for example, in the case of technical problems or mispronunciation). Any repeated tokens that were neither mispronounced nor had technical problems associated with them were included in the database. This resulted in roughly 800 fricative tokens per speaker, and roughly 900 tokens per fricative phoneme. The word tokens were hand-labelled by the first author using WAVES+. The total number of fricatives in the database was 7250, of which 3604 were voiceless. The following experiments were performed on the voiceless tokens.

In order to answer the questions posed in the introduction, we sampled our data at 44.1 kHz in order to look at the spectrum up to 20 kHz. The spectra were obtained using two different methods: a) a 1024 point FFT (representing 25 ms in time) was taken at the mid-point of the fricative token on speech windowed by a 1024 point Hamming window, and b) 1024 point FFTs were calculated across the entire token with a Hamming window-width of 1024 points and a frame-shift of 512 points. The spectrum of the fricatives was the root mean square of the spectrum $S_{AVE}(f)$. This was calculated by:

$$S_{AVE}(f) = \frac{1}{M}\sqrt{\sum_{m=0}^{M-1} S_m(f)^2}$$

where M was the number of FFTs obtained per token (this varied between 6 and 14 depending on the length of the fricative token), and where $S_m(f)$ is the magnitude spectrum obtained on a 1024 point FFT on speech samples windowed by a 1024 point Hamming window. Each consecutive $S_m(f)$ was overlapped by 512 samples. $S_{AVE}(f)$ is the root mean square average of the magnitude spectrum over the fricative signal, and varied between 81-174 ms, depending on the length of the token. The resulting spectrum was then amplitude-normalised to make the maximum amplitude in the spectrum one unit. The maximum amplitude of the spectrum was expected to be the spectral peak of the fricative token.

Both pre-emphasised and un-pre-emphasised data were investigated in each case, since the pre-emphasis would boost the higher frequency components which we wanted to explore. For the

pre-emphasised data, the pre-emphasised signal was calculated from x[n] - 0.95 x[n-1], which produces an approximate 6dB / octave rise to the spectrum.

All data were speaker-normalised following Lobanov (1971). Data were banded using quasi-Bark bands 1-24 (based on Zwicker 1961 -- henceforth simply 'Bark'), which cover the spectral range from 0 - 17,054 Hz. The microphone response was flat up to 18 kHz, and there was little roll-off effect from the anti-aliasing filter.

For all experiments, data were classified on an open test. Speakers were divided into a training set (consisting of three females and two males, with a total of 1977 fricative tokens) and a testing set (consisting of one female and three males, with a total of 1627 fricative tokens). Canonical Discriminant Analyses (CDAs) were performed on the training data for each new experiment. The CDA serves to reduce the number of dimensions, on which the testing data is classified, from 24 Bark bands to three transformed dimensions (the number of different fricative types, minus one -- Harrington and Cassidy, in press). The transformed testing data were classified using a Bayesian distance measure between the testing tokens and the training data.

Figure 1a shows an ellipse plot of the first two transformed dimensions from the CDA for the pre-emphasised training data. It can be seen that [s] is almost entirely separated from [S] on the first two transformed dimensions, but that there is almost total overlap between the non-sibilants. Figure 1b shows a normal distribution plot for the third transformed dimension: it can be seen that [f] is at least partly distinguished from [T] on this dimension.

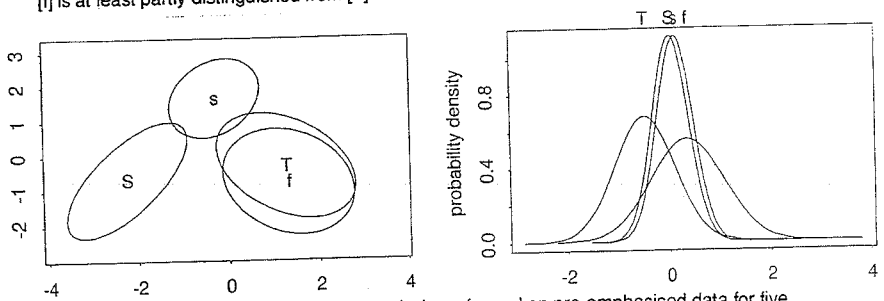

Figure 1: Canonical Discriminant Analysis performed on pre-emphasised data for five speakers (two male, three female). Data had been banded into quasi-Bark bands 1-24 (0 - 17,054 Hz).
a) ellipse plot for the first two transformed dimensions; b) normal-distribution plot for the third transformed dimension.

RESULTS

We present our results in answer to each of the three questions posed in the introduction. We should point out at this stage that results for [s] and [S] were always quite excellent. In all of our experiments, [s] was classified correctly approximately 95% of the time, and [S] was classified correctly approximately 93% of the time. Presentation of results will therefore focus on [f] and [T], and it will be seen that results were not as good for these non-sibilants as they were for the other two fricatives.

i) We found considerable differences in the number of [T] tokens correctly classified, according to whether we used single spectral slices taken at the midpoint, or averaged FFTs taken across the entire token (henceforth 'midpoint' and 'averaged' respectively). The averaged data always gave better results. There were no significant differences for the other fricatives. This was true of data

that were classified after doing the CDA, as well as of data that were classified before doing the CDA (i.e. classified on all 24 Bark bands).

Table 1 presents the confusion matrices for the spectral data taken at the midpoint, and the averaged spectral data. Training and testing tokens came from different speakers (as described above). The confusion matrix on the left is for averaged FFT data, and the matrix on the right for midpoint FFT data. Both are for pre-emphasised data which has been banded into Bark 1-24, and on which CDAs have been carried out. The results marked with double asterisks are significantly different at $p<0.01$, based on a paired t-test.

AVERAGED					MIDPOINT			
	S	s	T	f	S	s	T	f
S	93.5	6.2	0.0	0.2	93.5	6.5	0.0	0.0
s	3.7	95.8	0.2	0.2	2.6	97.0	0.2	0.2
T	0.3	0.8	84.1**	14.9	0.3	4.3	70.7**	24.7
f	0.5	0.0	35.2	64.3	0.0	0.7	31.9	67.3

TABLE 1 - Confusion matrices for averaged spectral data and for spectra taken at the midpoint. CDAs were performed on both sets of data. Results are for percentage correct classifications, and results marked with a double asterisk are significantly different at $p<0.01$, based on a paired t-test.

All results presented henceforth will be based on the averaged FFT data.

ii) We next compared pre-emphasised data with un-pre-emphasised data. We found no significant differences in the results between the two types of data. The only difference between the two was for [S], with 93% correctly classified in the case of pre-emphasised data, and 99% correctly classified in the case of un-pre-emphasised data. However, this difference was not significant based on a paired t-test. Rather than presenting both sets of data here, we have chosen to present only the results for the pre-emphasised data, since this seemed to provide a more consistent picture of [f] and [T] across the different experimental settings.

iii) We finally sought to find out whether the spectral information above 10 kHz was useful. We did this by reducing the number of Bark bands used in banding the training data and the testing data. By excluding Bark 24, we reduced the spectral range from 0 - 17,054 Hz to 0 - 12,618 Hz. By excluding Bark 23, we reduced the range still further to 0 - 10,164 Hz. The total number of correct classifications fell from 84% to 81% when Bark 24 was left out, but there was no difference in the overall number of correct classifications when Bark 23 was left out. The results for [f] in the case where CDA was not used and data were therefore classified on the 23 Bark bands was also significantly worse than in the case where all 24 Bark bands were used without the CDA.

Table 2 shows three confusion matrices. The first matrix gives results for all 24 Bark bands, the second for 23 Bark bands, and the third for 22 Bark bands. Training and testing tokens came from different speakers. All three matrices are for pre-emphasised data on which CDAs have been carried out. The results marked with an asterisk are significantly different at $p<0.05$, based on a paired t-test. In the matrices here as well as those in Table 1, it can be seen that [f] and [T] tend to be confused with each other, rather than with one of the other two fricatives.

BARK 1-24 (0-17,054 Hz)

	S	s	T	f
S	93.5	6.2	0.0	0.2
s	3.7	95.8	0.2	0.2
T	0.3	0.8	84.1	14.9
f	0.5	0.0	35.2	64.3*

BARK 1-23 (0-12,618 Hz)

S	s	T	f
93.3	6.5	0.0	0.2
4.0	95.3	0.2	0.5
0.0	5.1	81.6	13.4
1.0	0.5	42.9	55.6*

Bark 1-22 (0-10,164 Hz)

S	92.8	6.9	0.0	0.2
s	3.7	95.6	0.5	0.2
T	0.0	7.1	81.1	11.9
f	1.7	1.5	39.2	57.6

Table 2 - Confusion matrices for data using Bark 1-24 (0 to 17,054 Hz), 1-23 (0 to 12,618 Hz) and 1-22 (0 to 10,164 Hz). Results are for percentage correct classifications, and results marked with an asterisk are significantly different at $p<0.05$, based on a paired t-test.

It would therefore appear that there is some information for [f] above 12.5 kHz, which is consistent with Shadle (1985), whose calculations predicted a peak at 13 kHz for a labio-dental-like model. However, we do not feel that a clear picture is emerging here. For instance, when we attempted to classify the data using only the male speakers (training on four speakers and testing on the fifth male speaker), without doing CDA, the results were not significantly different with regard to [f] or [T]. We are unable to explain this, beyond a consideration of two points: 1) only male speakers were used, rather than a mix of female and male, which could explain the difference in frequencies, and 2) the CDA, as was shown above, goes some way in separating [T] from [f].

DISCUSSION AND FURTHER WORK

Our results show that the averaged spectral data provides much better results for [T] than the spectral slice taken at the midpoint. This would suggest that there is perhaps important spectral information in those parts of the fricative that are excluded when a slice is taken at the midpoint. For instance, it may be the case that effects due to the neighbouring vowel occur at the edges of the fricative. The coarticulatory influence exerted by vowels on fricatives is less in the case of [T] than in the case of [f], due to the relative freedom of the tongue body in the latter articulation, so that elements of the vowel spectra could have some influence on the fricative spectra near the fricative-vowel boundary. Work in progress aims to explore the dynamic properties of the spectra, using a Discrete Cosine Transform; and kinematic studies will examine the movement of the tongue body, blade and tip (as well as lip and jaw movement) during the production of the fricative/vowel sequence.

Pre-emphasis did not seem to make a large difference to results. We believe that the slightly worse performance of [S] with pre-emphasised data is due to the fact that pre-emphasis not only boosts the higher frequencies, it suppresses the lower frequencies. Since the peak for [S] is located at around 2 - 4 kHz, lower than for the other fricatives, it is to be expected that this would be the only fricative to show this effect.

The question of whether or not there is contributing information above 10 kHz is not resolved. Although results are significant in some instances with regard to spectral energy above 12.5 kHz, visual inspection of the spectra does not seem to support these results. In fact, our main impression for [T] is one of great inter-speaker variability, which may be due to different articulatory strategies (i.e. interdental vs. apico-dental). All things considered, it would appear that

the marginally better results obtained when Bark 24 is included in the analysis, do not justify a doubling of the sampling rate.

Further work will look at the vowel transitions as a cue to the discrimination between [f] and [T], and kinematic work will try to explain the anomaly of two such contrasting articulatory strategies having so similar an acoustic output.

ACKNOWLEDGMENTS

We would like to thank Jonathan Harrington for all his ideas on and help with the experiments, and for comments on earlier versions of this paper. We would also like to thank Chris Calaghan for supervising the recordings, and all our speakers for taking part.

REFERENCES

Behrens, S. & Blumstein, S. (1988) "Acoustic characteristics of English voiceless fricatives: a descriptive analysis," Journal of Phonetics **16** 295-298.

Fant, G. (1960) Acoustic Theory of Speech Production The Hague: Mouton.

Flanagan, J. (1972) Speech Synthesis, Analysis and Perception New York: Springer-Verlag.

Gurlekian, J. A. (1981) "Recognition of the Spanish fricatives [s] and [f]," Journal of the Acoustical Society of America **70** 1624-1627.

Lobanov, B. (1971) "Classification of Russian vowels spoken by different speakers," Journal of the Acoustical Society of America **49** 606-608.

Harrington, J. & Cassidy, S. (in press) Techniques in Speech Acoustics

Hughes, G.W. & Halle, M. (1956) "Spectral properties of fricative consonants," Journal of the Acoustical Society of America **28** 303-310.

Shadle, C. H. (1985) "The acoustics of fricative consonants," Ph.D. thesis, MIT.

Strevens, P. (1960) "Spectra of fricative noise in human speech," Language and Speech **3** 32-49.

Zwicker, E. (1961) "Subdivision of the audible frequency range into critical bands," Journal of the Acoustical Society of America **33** 248.