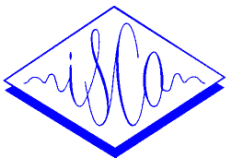

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

ENSEMBLE AVERAGING APPLIED TO THE ANALYSIS OF FRICATIVE CONSONANTS

Christine H. Shadle[1], André Moulinier[1], Christian U. Dobelke[2], Celia Scully[3]

[1]Department of Electronics and Computer Science
University of Southampton, Southampton SO9 5NH, England

[2]Preussen Elektra, Vor dem Nordwald 14, 3160 Lehrte, Germany

[3]Department of Psychology, University of Leeds, Leeds LS2 9JT, England

ABSTRACT

In an attempt to explain across-subject variability in fricative spectra, ensemble-averaged spectra were compared to time-averaged spectra for a corpus that consisted of nonsense syllables of the form /pV$_1$FV$_2$/ repeated 10-14 times on one breath. V$_1$ and V$_2$ were chosen from /a,i,u/; F was one of /f,v,θ,ð, s,z,ʃ,ʒ/. Two subjects were recorded saying both this corpus and another corpus consisting of the same fricatives sustained for 3 seconds. A wide variety of articulatory data was also available for these subjects. An ensemble-averaged spectrum could be computed for the beginning (or middle or end) of all fricative tokens. Results show a pattern of spectral change through the fricative that is consistent with aerodynamic and articulatory measures. The non-stationarity thus revealed does not in itself explain the variability across subjects, however. Rather, the ensemble averaging allows a precision in timing of the analysis window which, when coupled with the articulatory data, shows more clearly the effect of vowel context on fricatives, and delineates the differences between /f/ and /θ/.

INTRODUCTION

It is difficult to find distinguishing acoustic features for fricatives that are valid across subjects [1,2]. In a recent look at this problem, a corpus of sustained fricatives was examined for token-to-token variability within subject and across two subjects [3]. Token-to-token variability within subjects was surprisingly small, but significant across-subject variability was found, particularly for [ʃ]. Though this was shown to be consistent with articulatory differences, there was no apparent reason for either the articulatory or acoustic differences to exist.

For vowel-fricative-vowel (VFV) sequences the situation is complicated by the fact that the fricatives must be assumed to be non-stationary. Since the typical analysis technique of time averaging throughout the fricative is based on the assumption of stationarity, it seemed possible that the apparent variability across subjects might be due to a blurring introduced by the time averaging. Since fricatives are inherently noisy, some averaging must be done in order to observe the broad spectral characteristics. Ensemble averaging fulfills this requirement while not assuming the signal to be stationary [4]. Instead, an ensemble is required; though not present in ordinary speech, the corpus described below included an ensemble, and thus allowed both time- and ensemble-averaging to be performed.

This use of ensemble averaging was described recently [5] for a single subject. In this paper we include results for a second subject, and the VFV corpus analysis is compared to the analysis of sustained fricatives in [3] to ascertain: 1) the degree of difference between ensemble and time averaging, and thus the extent to which time averaging might cause apparent variability; 2) the way in which the spectral shape changes throughout the fricative; 3) the differences between /f/ and /θ/ spectra; 4) the effect of vowel context on the mid-ensemble-averaged spectra.

METHOD

Recording Method

The corpuses 2 and 3 used in this paper are part of a larger set of corpuses developed jointly at the University of Leeds, University of Southampton, and the Institut Communication Parlée, I.N.C.P., Grenoble. Corpus 2 consists of the fricatives /f,v,θ,ð,s,z,ʃ,ʒ,ç,j,x,ɣ,h/ sustained for 3 s. Six tokens of each fricative were recorded; the six sets of tokens were recorded in randomized order. In Corpus 3, the nonsense word /pV$_1$FV$_2$/ was repeated 10 to 13 times during a single breath. This set of repetitions was called an item. For each item, V$_1$ and V$_2$ were chosen from /a,i,u/; F was one of the set given above. Within each item, the first two and last two tokens of /pV$_1$FV$_2$/ were discarded, leaving 6 to 9 relatively uniform tokens forming the ensemble. Two speakers, a woman speaker of General American English and a man speaker of French (hereafter referred to as CS and PB respectively), were recorded speaking this corpus while a variety of acoustic, articulatory and aerodynamic measurements were made. For this paper, only the English fricatives /f,v,θ,ð,s,z,ʃ,ʒ/ were analyzed.

The main results shown here were based on spectral analysis of the high-fidelity acoustic recordings, which were made in an anechoic chamber using a Bruel & Kjaer 4165 1/2" microphone located 1 m in front of the subject's mouth. Recordings were made with a Sony PCM system at 16 bits with a sampling frequency of 44.1 kHz. A calibration signal was recorded to allow absolute pressure level to be retained. These results were also compared with separately obtained recordings using a Rothenberg mask and oral pressure measurement, and electropalatography (details are given in [6]).

Analysis Method

The analysis method differs for the two corpuses and is critical to the information obtained. For Corpus 2, where the fricative was sustained for 3s, an averaged power spectrum was computed by time-averaging 25 consecutive 20ms Hanning windows centered in the fricative. Since each fricative was said six times, six time-averaged power spectra resulted. For Corpus 3, each /pV$_1$FV$_2$/ combination was repeated 10 to 14 times within an item. These tokens were then analyzed in two ways: by time-averaging through the stationary part of the fricative in one token (and thus producing one time-averaged spectrum per token of an item), and by ensemble-averaging across the ensemble of tokens at specific events throughout the vowel-fricative-vowel sequence (and thus producing one ensemble-averaged

10.21437/ICSLP.1992-16

spectrum per event of an item). The ensemble averaging allows observation of changes in the spectrum over relatively short periods of time; it assumes that all tokens are produced under identical conditions. Time averaging allows observation of the way the spectrum varies for different tokens; it assumes that the portion of the fricative being averaged is stationary.

Both types of averaging depend on accurate labelling of events in the time waveform. Four points were labelled for each fricative: beginning of the vowel-fricative transition, beginning of fricative steady-state, end of fricative steady-state, and end of fricative-vowel transition. These four points establish the duration of the three phases of the /VFV/ transition: the vowel-fricative transition, the steady-state, and the fricative-vowel transition.

The start of the vowel-fricative transition phase is the same in both voiced and unvoiced cases. At a quite distinctive point the ampitude of the envelope of the vowel's fundamental frequency begins to decrease; this is assumed to be the start of this phase. During this phase, noise begins to add to the voicing, its extent dependent on the fricative.

The steady-state phase begins in unvoiced fricatives when the voicing of the preceding vowel has completely died out. For voiced fricatives, this phase is characterized in most cases by a constant amplitude of the fundamental frequency component.

The start of the fricative-vowel transition phase is characterized for unvoiced fricatives by the voice onset; the frication noise in the signal has completely disappeared. For voiced fricatives, this phase begins when the amplitude envelope of the fundamental frequency component begins to increase. The end of this phase is reached when the following vowel is in a steady state again, i.e. when the fundamental frequency component of this vowel remains at a constant amplitude.

Special cases were considered separately and resulted at times in the labelling of more points, e.g. to mark a region of devoicing in the center of a voiced fricative.

The labelled points in each token were then used to locate analysis windows, as shown in Fig. 1. For the ensemble averaging, three locations were routinely used: the 20ms beginning at the label marking the beginning of the fricative steady-state (referred to as *beg*), the 20ms centred between *beg* and *end* (where *end* is the label marking the end of the fricative steady-state), and the last 20ms before *end*. Subject PB always produced 10 tokens per item for Corpus 3; of these, the first and the last were ignored, and ensemble averages were performed over the 8 tokens numbering 2 through 9. Subject CS produced a variable number of tokens until she ran out of breath; for her, ensemble averages were performed over the 8 tokens numbering 2 through 9.

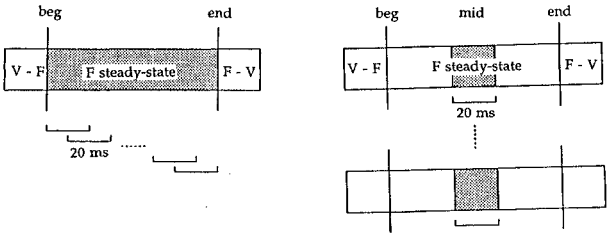

Figure 1: Diagram showing the placement of labels and windows in a fricative, shown here subdivided into two transition regions (V-F and F-V) and a steady-state region. The shaded portions are included in one or more analysis windows. Left: time averaging uses overlapping windows within one token. Right: ensemble averaging uses one window per token, here, centered in each fricative.

RESULTS

Corpus 2

These results were reported in Ref. [3]. A summary of the principal findings follows.

The time-averaged spectra derived from each token are shown in [3] with all six tokens for each fricative on one graph. The amount of variability for each fricative can thus be immediately assessed. The productions look quite consistent in overall level, general spectral shape, and even in many details of the spectral shape, that is, in formant location. This observation held even for [ç], which was non-native for both subjects.

Using the results of the transfer function measurements, formants and their approximate cavity affiliations were identified. The particular formants making up a broad peak were not always the same for both subjects; in particular, the frequency-ordering of zeros and poles differed, which could have a substantial effect on the broad spectral shape. For [ʃ], a case where such differences were striking, it was found that the spectral differences could be attributed to a difference in the frequency of the free zero, which was consistent with articulatory data (direct palatography). Essentially, PB made the constriction further back than did CS. It was not known to what this consistent difference could be attributed at the time of presenting the paper, but a plausible explanation has occurred in light of the Corpus 3 findings, as detailed below.

Corpus 3

The segmentation rules developed in order to locate the labels used in the analysis were verified by two means: the amount of variation in duration of the three phases of the fricatives was checked for each item, and these durations were checked where possible against durations reported for segmentation performed using the same subjects but other input signals (e.g. EPG, airflow, or oral pressure). The variations within an item were small, leading one to the conclusion that both production and segmentation were fairly uniform. Some general patterns across items appeared consistent, also: the steady-state portion was longer for unvoiced than for voiced fricatives, and it occupied a greater proportion of the total fricative length. The durations measured on other signals were not always within a standard deviation of the mean durations measured here, but were consistently larger or smaller. Thus although different segmentation criteria were apparently used, these events so identified consistently occurred in the same sequence. This allows events observable in different sets of measurements to be interrelated.

Having verified the labelling, several comparisons were made, which we will consider in turn. First, time- and ensemble-averaged spectra were compared. Figure 2 contrasts two cases for subject CS. The great similarity evidenced in /pasa/ holds for nearly all fricatives in both subjects. The spectra for /piθi/ show one of the exceptional cases. Here some of the tokens are considerably weaker, causing the respective time averages to differ considerably. Clearly the assumption that all tokens are produced under identical circumstances is a poor one in such a case. Ideally, tokens should be screened for similarity by using time-averaged spectra, and then only similar tokens should be ensemble-averaged to explore spectral changes at specific points in the fricative.

Second, the ensemble-averaged spectra computed at three points during the fricative steady-state were compared within each item. The three points at which the analysis windows were located were: starting at *beg*, centered in the steady-state region, and ending at *end*. In the discussion that follows these spectra

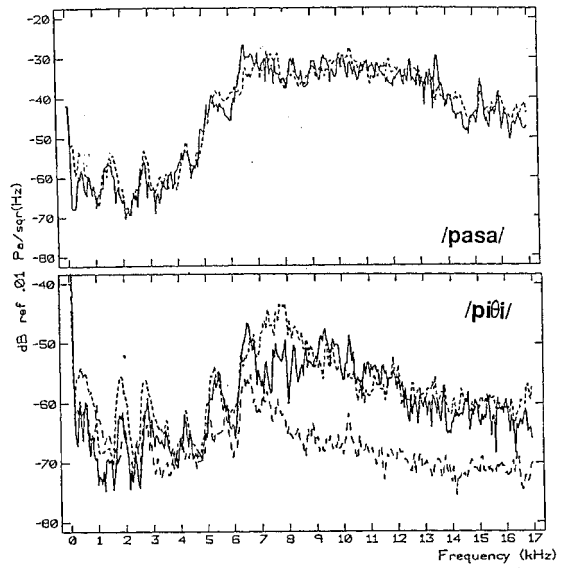

Figure 2: Comparison of time- and ensemble-averaged spectra for subject CS. Time-averaged spectra (dotted and dashed lines) and ensemble-averaged spectra at fricative centers (solid line) are shown for (a) /pasa/ (b) /piθi/.

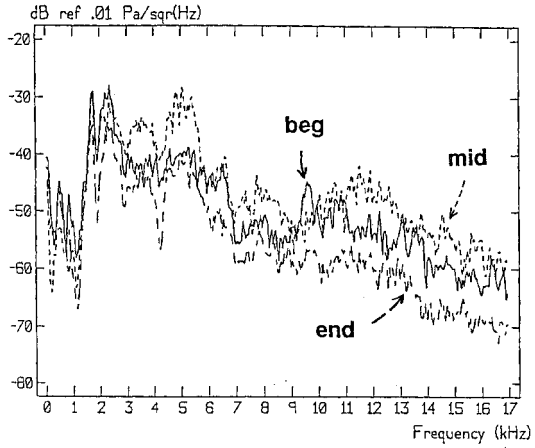

Figure 3: Three ensemble-averaged spectra from /paʃa/, subject PB. Solid line is beg-spectrum, dotted line is mid-spectrum, dashed line is end-spectrum.

will be referred to as the beg-spectrum, the mid-spectrum, and the end-spectrum. Figure 3 shows an example of these three spectra from /paʃa/ for PB; Fig. 2 of Ref. [5] shows an extended example (including two spectra from the transition regions) from /pasa/ for CS.

In nearly all cases the ensemble-averaged spectra computed at these three points in the steady-state region are significantly different; further, the differences generally follow the same pattern. The mid-spectrum tends to have the highest amplitude at high frequencies and the lowest amplitude at low frequencies. The crossover point depends on the particular fricative, but is approximately 5 kHz or less. The formant peaks visible at lower frequencies are usually nearly the same in the three spectra, but the noise levels at higher frequencies can differ by 10 dB or more. Typically at the highest frequencies examined, i.e. near 17 kHz, the mid-spectrum has the highest amplitude, the beg-spectrum the next highest, and the end-spectrum the lowest.

Comparison with articulatory and airflow data indicates

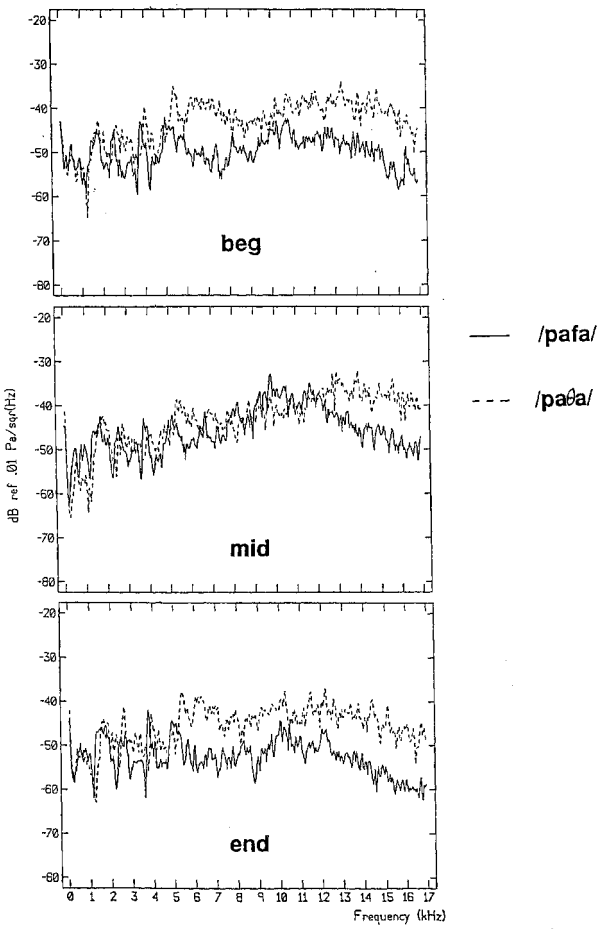

Figure 4: Ensemble-averaged spectra contrasted for /pafa/ (solid line) and /paθa/ (dotted line), subject PB. Top: beg-spectra. Centre: mid-spectra. Bottom: end-spectra.

that there are three major adjustments that must be made in the transition region. The airflow increases; the vocal folds are abducted, and they stop vibrating; the tongue must move to form the constriction. The latter takes the longest, so as a result frication begins while the constriction area is still decreasing. As the area grows smaller the air velocity through the constriction increases, which boosts the high-frequency energy and simultaneously cancels low-frequency back cavity formants more effectively [6,7]. For /s, ʃ/ different vowel contexts affect the formant trajectories into and out of the fricative, but do not substantially alter the sequence or timing of events.

For the weak fricatives, however, the differences are more marked. /pafa/ and /paθa/ are contrasted for both subjects: Fig. 4 shows data for PB, and data for CS is given in Fig. 3 of Ref. [5]. In both cases, the mid-spectra are similar, but the beg- and end-spectra differ. Above 7 kHz for CS, 5 kHz for PB, /paθa/ has a higher amplitude than /pafa/ at the edges of the steady-state region. Its formants also increase in frequency relative to those of /pafa/, supporting the generally accepted view [8] that the difference between these fricatives lies in the formant transitions.

In the voiced fricatives, the fundamental frequency peak remains of high amplitude throughout the fricative, but the amplitude at low frequencies (above F0 and below approximately 5 kHz) does drop generally, consistent with unvoiced fricatives. The amplitude above 5 kHz is lower than the unvoiced counter-

part, but the basic shape and the pattern of the beg-, mid- and end-spectra are generally similar.

The effect of vowel context was also studied, mainly by comparing mid-spectra. For all fricatives, the /a-a/ and /i-i/ contexts were similar at high frequencies in broad spectral shape and level. Formant frequencies were sometimes higher for the /i-i/ context. The /u-u/ context was noticeably different for every fricative: formant frequencies were lowest, overall level was lowest, and in addition the broad spectral shape could differ significantly. The most striking example of this is shown in Fig. 5, where mid-spectra for /pasa/ and /pusu/ are contrasted. It is well known that the lip-rounding for /u/ lowers resonance frequencies and decreases overall amplitude, but these changes cannot explain all of the differences between these spectra. By comparison with the changes occuring within a fricative, it appears that the /s/ in /pusu/ has been produced with a much lower airflow, resulting in a weaker and less broad-spectrum noise source.

Finally, the mid-fricative ensemble averages were compared to the Corpus 2 spectra. It was expected at the outset that the sustained fricatives in Corpus 2 were produced with the most stable articulatory configuration, and were in a sense the 'best' version of each fricative, at which the unsustained fricatives in Corpus 3 were aimed. The degree to which the mid-spectrum of a Corpus 3 item matched the Corpus 2 tokens of that fricatives would then presumably depend on how much time the tongue had to get into the fricative configuration. The further the tongue had to travel between vowel and fricative, the more dissimilar the sustained and unsustained fricative spectra were expected to be. Thus /piθi/ should match the Corpus 2 /θ/ better than /paθa/, /pisi/ should match /s/ better than /pasa/, and so on.

In general at least one of the items for a fricative in Corpus 3 matched within the variation defined by the six tokens of Corpus 2. Of the three vowel contexts /a-a/, /i-i/, /u-u/, for PB the /aFa/ item always matched the Corpus 2 spectra best. For CS, usually the /aFa/ item matched best, but sometimes /iFi/ did. This difference can best be explained by the subjects' difference in production of Corpus 2: PB always preceded the sustained fricative by /a/; CS instead assumed the fricative from a silent, neutral position. This may also explain the difference in constriction location for [ʃ] noted in the Corpus 2 analysis: PB used a more posterior constriction than CS, consistent with the difference in vowel context.

Since the /a-a/ context matched the most often, contradicting the prediction, the idea of a single target for each fricative must be abandoned. It appears instead that vowel context may influence even a sustained fricative, and since the context can change spectral shape significantly, these changes must be understood to color the fricative but not change its identity. It thus remains unclear exactly which features do allow us to distinguish fricatives, but the vowel-fricative transition regions provide important clues which can be examined with precision using the technique of ensemble averaging.

CONCLUSION

Both time- and ensemble-averaged spectra were computed for repeated /pV$_1$FV$_2$/ tokens for two subjects in order to investigate causes of across-speaker variability, focus on the details of spectral change throughout a fricative, and establish the effects of the vowel context. Ensemble averages computed in the center of the ensemble of fricatives were in general quite similar to time averages computed throughout single

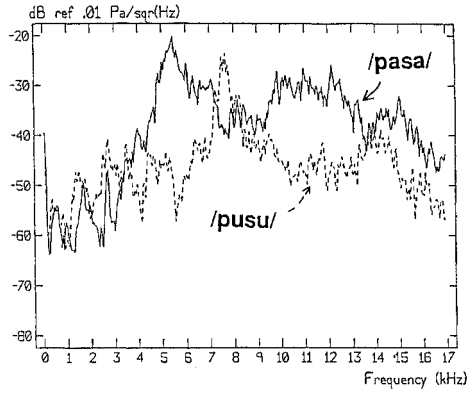

Figure 5: Ensemble-averaged spectra from the centers of /pasa/ (solid line) and /pusu/ (dotted line), subject PB.

fricatives. However, ensemble averages at beginning and end of a fricative did differ significantly from the central averages, and in accordance with the articulatory and aerodynamic changes occurring in the transition regions. Vowel context was shown to have a significant effect on the mid-ensemble-averaged spectra and on sustained fricatives, explaining apparent cross-subject variability for sustained /ʃ/. In summary, the use of time averaging does not introduce variability even though it includes the transition regions; rather, ensemble averaging allows a more detailed dissection of the many influences on the spectral shape of fricatives.

Acknowledgements

This work was supported in part by a collaborative EC SCIENCE award, CEC-SCI*0147C(EDB), and by a joint CNRS/Royal Society award.

References

1. W. Hughes and M. Halle. 'Spectral Properties of Fricative Consonants,' *J. Acoust. Soc. Am.* 28, 303-311, 1956.

2. J.M. Heinz and K.N. Stevens. 'On the Properties of Voiceless Fricative Consonants.' *J. Acoust. Soc. Am.* 33, 589-596, 1961.

3. C.H. Shadle, P. Badin, and A. Moulinier. 'Towards the Spectral Characteristics of Fricative Consonants,' Proc. of the XIIth Int. Cong. of Phonetic Sciences, Aix-en-Provence, v.3, 42-45, 1991.

4. J.S. Bendat and A.G. Piersol. *Random Data*, Wiley-Interscience, 1971.

5. C.H. Shadle, C.U. Dobelke, and C. Scully. 'Spectral Analysis of Fricatives in Vowel Context,' *J. de Physique IV*, Coll. C1, supp. *J. de Physique III*, v.2, C1-295 to C1-298, 1992.

6. C. Scully, E. Georges, and E. Castelli. 'Fricative Consonants and their Articulatory Trajectories,' Proc. of the XIIth Int. Cong. of Phonetic Sciences, Aix-en- Provence, vol.3, 58-61, 1991.

7. C.H. Shadle. 'Articulatory-Acoustic Relationships in Fricative Consonants,' in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal (eds.), Kluwer, 187-209, 1990.

8. K. Harris. 'Cues for the Discrimination of American English Fricatives in Spoken Syllables,' Language and Speech 1, 1-7, 1958.