SPEAKER IDENTIFICATION USING A SPECTRAL MOMENTS METRIC WITH THE VOICELESS FRICATIVE /s/

R-M. de Figueiredo and S-L. Olivier
Department of Legal Medicine-FCM-UNICAMP, Campinas, Brazil

ABSTRACT

FFT spectra of 26 voiceless fricatives /s/ (10 speakers) were treated as a random probability distribution from which the first four moments were computed. In the first experiment a discriminant analysis based on the four moments resulted in correct classification of speaker identity ranging from 60% to 90%. In a second experiment, a cross-validation test showed that new samples may be correctly matched with the reference material in 8 cases out of 10.

INTRODUCTION

Fricative sounds have already been shown to carry relevant information concerning some of the speakers' characteristcs such as sex and identity [1, 2, 3, 4, 5]. However, the experiments conducted in the studies which proved that to be true concentrate on perceptual evaluation of fricatives produced in isolation and/or in the same phonetic context, neglecting the expected variation intra-speaker in the production of fluent speech. The present work aims at evaluating the efficacy of the fricative /s/ in identifying speakers, using speech samples extracted from fluent reading in various phonetic environments and at two diferent speech rates.

METHOD

Subjects

Ten male subjects aged between 24 and 42 were selected for study. the speakers were free from any speech defect and spoke general Brazilian Portuguese.

Materials and recording conditions

The speakers were asked to read a text extracted from a cientific journal in two different conditions: (a) at a normal and comfortable rate, and (b) as fast as possible, while mantaining intelligibility. Speech samples were all recorded analogically using a high-quality equipment (GRADIENTE Esotech DII tape recorder, and REALISTIC 33984-C microphone) in an acoustically isolated room with no specific reflection characteristics.

Fifty-two voiceless fricatives ocurring in various contexts were extracted from this basic material (twenty-six for each speech rate condition). Only fricatives in CV stressed syllables were used. V is one of the seven Brazilian Portuguese oral vowels (/a, ɛ, e, i, ɔ, o, u/). The words that were analysed in this study were not balanced for vowel context. It means that the number of cases in each vowel context is not, necessarily, the same.

Procedures

The CSL 4300B (KAY Elemetrics Corp.) was used for all acoustic analysis. The signal was digitized by 12-bit ADC at a sampling rate of 25 KHz. Following sampling, a digital high-pass filter with a 200 Hz cutoff frequency was applied to the speech waveform, in order to reduce low-frequency extraneous interference resulting from room vibration.

Only the median third of each fricative was selected for the extraction of each cross-sectional spectrum, in order to minimize any effects of anticipatory co-articulation with the neighbouring vowel. For each [s]-kernel a 512-point fast Fourier transform (FFT) was computed.

After normalized by peak, each FFT-spectrum (only range 0.5-10KHz) was treated as a random probability distribution from which the first four moments (mean, variance, skewness and kurtosis) were calculated. Figures 1 and 2 show pairs of spectra that differ in some of these values.

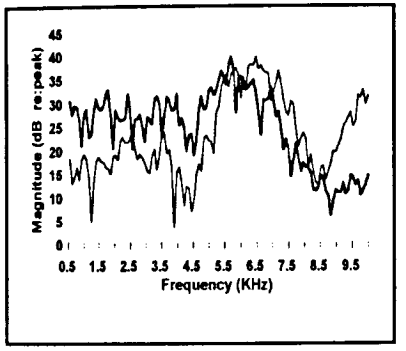

Figure 1: Two spectra, of different speakers, that differ in skewness and mean. The thin-lined spectrum has higher mean and slightly negative skewness, while the thick-lined spectrum has lower mean and markedly positive skewness.

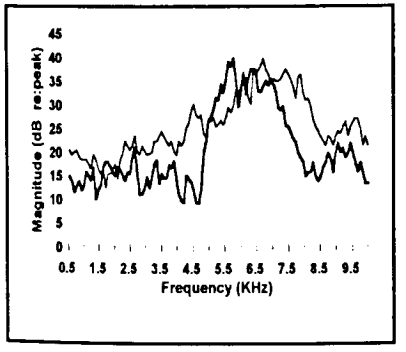

Figure 2: Two spectra, of different speakers, that are basically different for the value of the kurtosis (positive in the thick -lined spectrum and near zero in the thin-lined one)

The values of mean, standard deviation, skewness and kurtosis, derived from the cross-sectional spectra, served as input for a stepwise discriminant analysis accomplished with the program BMDP-7M [6]. The program finds the combination of variables that best predicts the group (speaker) to which a case (represented by the four moments) belongs. At each step, the variable that adds the most to the separation of the groups is entered into the discriminant function. In the end of the process, all variables that, in any way, contribute to the separation of speakers (according to a predetermined minimal F-value) enter the discriminant function.

At a first stage, the efficacy of this spectral moments metric was tested by using only the 26 fricatives produced at a normal speech rate. At a second stage, the results undergo a cross-validation test, in an attempt to classify the 26 fricatives produced under the fast speech rate condition, on the basis of the discriminant functions obtained in the first stage.

RESULTS

Table 1 shows a classification matrix obtained in the first stage of the experiment. The basis for the results found at this point was only the 26 [s]-kernels extracted from the speech samples at normal speech rate. Table 2 shows the results of the cross-validation test. At this point, the newset of 26 fricatives extracted from the fast speech was classified according to the discriminant function obtained in the first test. The observation of table 2 reveals that only two out of ten speakers were not correctly classified (S3-F and S6-F). It should also be noticed that the percentage of correctedness in general decreases considerably in relation to the first test (see table 1), in which only fricatives extracted from speech samples at normal speech rate were employed.

CONCLUSION

The results suggest that voiceless fricatives /s/ in CV stressed syllables are, potentially, good indicators of the

identity of the speaker, even if extracted from fluent speech and in different vowel contexts. Nevertheless, due to expressive alterations in the speed of production, the percentage of correct classification decreases considerably.

It is also important to observe that the efficacy of fricatives [s] in identifying speakers is doubtful in the forensic paradigm, in which the recording quality and bandwidth, both present in this experiment, should not be expected. On the other hand, in speaker automatic verification systems, in which it is possible to control a series of conditions (background noise, sound quality, etc) the use of fricatives seems to be potentially interesting.

Table 1. Classification matrix showing the percentage of cases classified in each group (speaker). The cells in boldface show the percentage of correct classifications.

Subject	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	**92.3**	0.0	0.0	0.0	3.8	0.0	0.0	0.0	3.8	0.0
S2	0.0	**65.4**	7.7	15.4	0.0	0.0	0.0	7.7	3.8	0.0
S3	3.8	7.7	**76.9**	3.8	0.0	3.8	0.0	0.0	0.0	3.8
S4	3.8	7.7	11.5	**61.5**	3.8	0.0	0.0	7.7	0.0	3.8
S5	0.0	0.0	0.0	0.0	**80.8**	0.0	7.7	0.0	11.5	0.0
S6	0.0	0.0	3.8	3.8	0.0	**73.1**	0.0	0.0	0.0	19.2
S7	0.0	7.7	3.8	0.0	0.0	0.0	**88.5**	0.0	0.0	0.0
S8	0.0	3.8	0.0	7.7	0.0	7.7	0.0	**73.1**	7.7	0.0
S9	0.0	0.0	0.0	0.0	7.7	0.0	0.0	0.0	**76.9**	15.4
S10	0.0	0.0	15.4	3.8	0.0	11.5	0.0	0.0	7.7	**61.5**

Table 2. Results of the cross-validation test showing the percentage of classifications of the test group (only samples of fast speech: S1-F, S2-F, etc) in relation to the reference group, based on the discriminant functions obtained in the first phase. The cells in boldface highlight the higher percentage on each line.

Subject	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1-F	**61.5**	11.5	0.0	0.0	15.4	0.0	0.0	0.0	7.7	3.8
S2-F	0.0	**34.6**	0.0	26.9	0.0	23.1	0.0	3.8	11.5	0.0
S3-F	**42.3**	3.8	38.5	0.0	0.0	11.5	0.0	0.0	0.0	0.0
S4-F	15.4	23.1	0.0	**53.8**	0.0	0.0	0.0	7.7	0.0	0.0
S5-F	3.8	0.0	0.0	3.8	**57.7**	0.0	3.8	0.0	15.4	7.7
S6-F	0.0	**46.1**	0.0	0.0	0.0	30.8	0.0	19.2	0.0	3.8
S7-F	0.0	3.8	7.7	0.0	7.7	0.0	**80.8**	0.0	0.0	0.0
S8-F	19.2	26.9	0.0	3.8	0.0	7.7	0.0	**42.3**	0.0	0.0
S9-F	0.0	11.5	11.5	0.0	7.7	0.0	0.0	0.0	**46.1**	23.1
S10-F	0.0	0.0	7.7	11.5	0.0	19.2	0.0	0.0	15.4	**46.1**

REFERENCES

[1] Ingemann, F. (1968), "Identification of the speaker's sex from voiceless fricatives, *JASA* 44, 1142-44

[2] Schwartz, M. (1968), Identification of speaker sex from isolated, voiceless fricatives", *JASA* 43, 1178-1179

[3] La Riviére, C. (1974), "Speaker identification from turbulent portions of fricatives", *Phonetica* 29, 246-252

[4] Wu, K. and D.G. Childers (1991), "Gender recognition from speech. Part I: Coarse analysis", *JASA* 90, 1828-1840

[5] Hirson, A. and M. Duckworth (1993), "Glottal fry and voice disguise: a case study in forensic phonetics", *J. Biomed. Eng* 15, 193-200

[6] Jennrich, R. and P. Sampson (1990), "7M: Stepwise discriminant analysis", *BMDP Statistical Software Manual*, Univ. Calif. Press, 339-358