AN ACOUSTIC-PHONETIC FEATURE-BASED SYSTEM FOR THE AUTOMATIC RECOGNITION OF FRICATIVE CONSONANTS

Ahmed M. Abdelatty Ali [(1)], *Jan Van der Spiegel* [(1)] *and Paul Mueller*[(2)]

[(1)] Department of Electrical Engineering,
University of Pennsylvania,
Philadelphia, PA 19104-6390

[(2)] Corticon, Inc.
155 Hughes Rd,
King of Prussia, PA 19406

ABSTRACT

In this paper, the acoustic-phonetic characteristics and the automatic recognition of the American English fricatives are investigated. The acoustic features that exist in the literature are evaluated and new features are proposed. To test the value of the extracted features, a knowledge-based acoustic-phonetic system for the automatic recognition of fricatives, in speaker independent continuous speech, is proposed. The system uses an auditory-based front-end processing and incorporates new algorithms for the extraction and manipulation of the acoustic-phonetic features that proved to be rich in their information content. Several features, which describe the relative amplitude, location of the most dominant peak, spectral shape and duration of unvoiced portion, are combined in the recognition process. Recognition accuracy of 95% for voicing detection and 93% for place of articulation detection are obtained for TIMIT database continuous speech of 22 speakers from 5 different dialect regions.

1. INTRODUCTION

The fricatives form the largest set of consonants in the English language which has nine standard fricative consonants, namely: the voiceless fricatives which include the labio-dental /f/ as in leaf, the linguo-dental /th/ as in teeth, the alveolar /s/ as in lease and the palatal /sh/ as in leash and their voiced cognates /v/ as in leave, /dh/ as in seethe, /z/ as in Lee's and /zh/ as in rouge. The ninth fricative is the /h/ which is considered also a semivowel. These consonants can be distinguished by English-speaking listeners in identical phonetic contexts, regardless of whether these contexts are meaningful utterances or nonsense syllables. Therefore, the features needed for such discrimination can only reside in the acoustical signal.

Several past studies have investigated such features. References [9, 11, 12] are examples of some of the earliest studies which characterize the fricative consonants. Using perceptual experiments on synthetic speech, analysis of spoken syllables and primitive recognition experiments, these studies provide us with much data on the acoustic characteristics of fricatives. However they have been largely qualitative in nature and relied on a small set of stimuli produced in few vowel contexts usually by a single speaker. Later studies [6, 8, 15, 16] have added to our knowledge about fricatives. However, except for a few studies, the acoustic characteristics that exist in the literature are qualitative, relational and speaker dependent. They characterize the fricatives well from the articulation standpoint of separate syllables. When it comes to the automatic recognition of continuous, naturally spoken, speech, a considerable amount of research is still needed. Some recent studies [2, 3, 4, 10, 13, 14, 18] have tried to deal with this problem but more work still needs to be done until we are able to fully understand the variability of the acoustic characteristics of the fricative consonants.

In this paper, we discuss the results of our research in this area. The /h/ phoneme is excluded from our experiments due to its semi-vowel unique characteristics. The literature's acoustic-phonetic features of fricatives are tested carefully for their information content using different methods of extraction. New features are proposed and tested as well. Eventually, a complete fricative recognition system is simulated which incorporates new algorithms for the extraction and manipulation of the information-rich features. For space reasons, only the final results are given. A more detailed discussion of the features involved, and their characteristics, is given in [1].

2. FRICATIVE RECOGNITION SYSTEM

2.1 Front End Processing

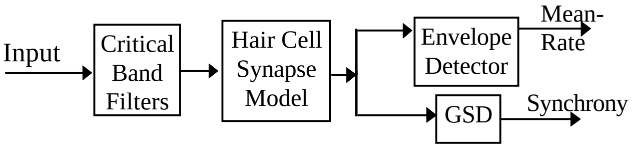


Fig.(1). Block diagram of an auditory-based front-end signal processing system.

The front-end signal processing that is used in our system is a biologically-oriented filter-bank system. It is based on the system developed by Seneff and described in detail in [17]. The block diagram is given in Fig.(1). The system gives two outputs, namely the mean-rate output and the Generalized Synchrony Detector (GSD) output. The front-end Bark-scaled filter bank consists of 36 filters with 20dB/decade high frequency pre-emphasis. The reasons for choosing this system are described in detail in [1].

2.2 The Recognition Experiments

The system mentioned in this paper is designed using 220 fricatives extracted from continuous speech from the TIMIT database spoken by 6 speakers, 3 females and 3 males having a northern dialect of American English. The system is then tested on 500 different fricatives extracted from the continuous speech TIMIT database for 22 speakers from 5 different American dialect regions, namely: northern, western, southern, midland and New York city dialects. The individual features are tested separately [1]. After deciding on which features to use, based on their information content, an automatic recognition algorithm was developed to combine the different features into a single decision. The results of this algorithm are described below.

2.3 Acoustic Features Used

2.3.1 *Duration and Voicing Detection*

The duration of the unvoiced portion (DUP) of the fricative is used as a voicing detection feature [1, 18]. We developed a method to detect the absence or presence of voicing in the signal and hence detect the start and end points. Voicing is manifested in the output by low frequency energy which is characteristic of voiced sounds, specially vowels and semi-vowels. Two methods are developed to detect such energy [1]:

1. The total energy of the lowest 9 filters (less than 1 kHz) in the GSD output. Such output is readily normalized in the GSD processing and is increasingly sensitive to periodicity. Call it LOWG.
2. The ratio between the low frequency (below 1.5 kHz) and high frequency (above 3 kHz) energies in the mean rate output, normalized with respect to the nearest vowel. Call it LOWE.

The advantage of using two quantities instead of one is that they tend to complement each other. If *either* of them exceeds its threshold, then phonation is assumed present. Therefore the DUP is the period where both quantities (LOWG and LOWE) are below their respective thresholds. Clearly, for a fully phonated fricative, DUP is equal to zero. If DUP is below a certain, empirically determined, threshold then the fricative is assumed voiced, otherwise it is voiceless.

	Detected as voiced	Detected as unvoiced
Voiced	186	17
Unvoiced	9	288

Table (1). Confusion Matrix for voicing detection Correct response rate is 95%.

The threshold used for the DUP is about 60 ms. However, if the DUP is above 100 ms, then, almost surely, the fricative is voiceless. This is in agreement with Stevens result, who found 60 ms to be a threshold for voiceless detection [18]. The advantage of using the two features LOWG and LOWE, instead of just one of them could be confirmed by comparing the performance of table (1) to those obtained by using either feature alone. The mean-rate quantity (LOWE) gives 85% correct response, while the GSD quantity (LOWG) gives 88%. The 85% obtained from the mean rate quantity (LOWE) is in agreement with Stevens result of 83%. The 10% improvement (from 85% to 95%) results from the use of the GSD output with its powerful periodicity detection ability [1, 17].

2.3.2 *Relative Amplitude and Spectral Flatness*

Relative amplitude (intensity) (RA) has been suggested in the literature as a feature to discriminate between sibilants (alveolars and velars) which have large RA and non-sibilants (labio- and linguo-dentals) which have small RA. The RA is defined as: $RA = \sum_{i:all\,filters} yenv_i|_{fricative} \Big/ \sum_{i:all\,filters} yenv_i|_{vowel}$

where $yenv_i$ is the mean-rate output from the ith filter and normalization takes place with respect to the nearest vowel.

A better performance is obtained by integrating two properties, namely: the low relative amplitude and the spectral flatness which characterizes non-sibilants, in one feature that could be used solely for the discrimination between sibilants and non-sibilants. We called this feature the Maximum Normalized Spectral Slope (MNSS). It is defined as:

$$MNSS = \max_{i:all\,filters}\left\{(yenv_i - yenv_{i-1})|_{fricative}\right\} \Big/ \sum_{i:all\,filters} yenv_i|_{vowel}$$

This feature gave excellent performance and explained the results obtained earlier by Behrens and Blumstein in their perceptual experiments [1,3]. A threshold was chosen empirically to be 0.02 for unvoiced and 0.01 for voiced fricatives. Thus, if MNSS is greater than 0.02 (or 0.01), the fricative is detected as a sibilant, otherwise it is non-sibilant. If MNSS is near the threshold value, we use normalization with respect to the fricative energy instead of the nearest vowel. The results are shown in table (2).

	Detected as sibilant	Detected as non-sibilant
/s/ and /z/	89	0
/f/, /v/, /th/ and /dh/	8	83
/sh/ and /zh/	34	6

Table (2). Confusion matrix for sibilant/non-sibilant discrimination using the MNSS. Correct rate is 94%.

2.3.3 *Spectral Shape and Peak Location*

The spectral shape is known to play a major role in the place of articulation detection of fricatives. Alveolar fricatives are characterized by a higher lowest spectral peak compared to palatal fricatives. Since labio- and linguo-dentals have been successfully detected in the previous section because of their relatively flat spectrum and low amplitude, the concern in this section will be on how to discriminate between alveolars and palatals.

The primary feature that we investigated is the most dominant peak (MDP) location. Palatals are characterized by a compact spectrum which has a dominant peak at a relatively low frequency, compared to the alveolars whose peak is at a higher frequency, and to the non-sibilants which usually do not have a significant peak and their most dominant peak is usually at a higher or much lower frequency. Therefore, this feature could be useful in extracting palatals. The best performance is obtained using the GSD output which yielded a better performance compared to the mean-rate (98.5% versus 91%) [1]. The results are shown in table (3).

	Detected as palatal	Detected as non-palatal
/s/ and /z/	0	89
/f/, /v/, /th/ and /dh/	0	91
/sh/ and /zh/	37	3

Table (3). Confusion matrix for palatal detection using the GSD MDP. Correct response rate is 98.5%.

Another feature which was found to play an auxiliary role in the discrimination between alveolars and palatals is the Spectral Center of Gravity (SCG). It describes some properties of the spectral shape which are not described by the MNSS or the MDP location. It is defined as:

$$SCG = \sum_{i:all\,filters>1.2\,kHz} i \times yenv_i \Bigg/ \sum_{i:all\,filters>1.2\,kHz} yenv_i$$

2.4 Place of Articulation Detection

In the previous section, the different acoustic features that are needed in the place of articulation detection are extracted and evaluated separately. In this section, the different features are combined to form a decision on the place of articulation. We use 3 main features:

- The Maximum Normalized Spectral Slope (MNSS).
- The Spectral Center of Gravity (SCG).
- The location of the most dominant peak (MDP).

	Detected as alveolar	Detected as dental	Detected as palatal
Alveolars:/s/ and /z/	85	4	0
Dentals: /f/, /v/, /th/, and /dh/	0	91	0
Palatals:/sh/ and /zh/	2	0	38

Table (4). Confusion matrix for place of articulation detection for the 6 speakers (3 males and 3 females) used in the system design. Correct response rate is 97%.

The algorithm used is explained in Fig.(2). The results of the place of articulation detection are represented in the confusion matrix of table (4) for the data used in the design and in table (5) for new data that were never encountered before by the system. The recognition rate obtained using this algorithm is about 97% for the former and 93% for the latter. For multi-speaker continuous speech recognition, this is a very good result given the simplicity of the algorithm and the system used.

	Detected as alveolar	Detected as dental	Detected as palatal
Alveolars: /s/ and /z/	188	11	5
Dentals: /f/, /v/, /th/, /dh/	2	144	6
Palatals: /sh/ and /zh/	8	2	134

Table (5). Confusion matrix for place of articulation detection for 22 new speakers, from 5 different dialects, *not* used in the system design. Correct rate is 93%. (alveolars: 92%, dentals: 95% and palatals: 93%).

2.5 Overall Recognition

The voicing detection and the place of articulation detection are combined in a single system which is capable of differentiating between fricatives. The results are shown in table (6).

	/s/	/f/, /th/	/sh/	/z/	/v/, /dh/	/zh/
/s/	90	8	2	0	0	0
/f/, /th/	4	87	4	0	5	0
/sh/	4	1	92	0	0	3
/z/	8	1	1	85	2	3
/v/, /dh/	0	0	0	0	100	0
/zh/	0	0	9	7	1	83

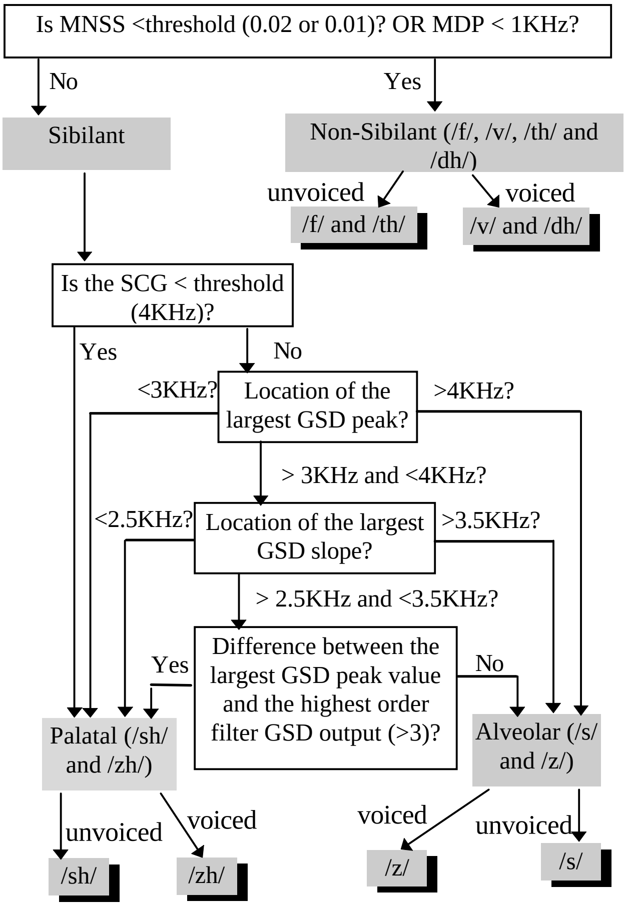
Table (6). Confusion matrix for fricatives' detection (in percentages) for 22 speakers, from 5 different dialects, *not* used in the system design. Overall correct recognition rate is 90%. Rows are inputs, columns are outputs.

3. CONCLUSION

In our work, the acoustic features characterizing the fricative consonants are analyzed in detail. Three features proved to be very useful in detecting the place of articulation, namely: the Maximum Normalized Spectral Slope (MNSS), the location of the Most Dominant Peak (MDP) and the Spectral Center of Gravity (SCG). These features were able to achieve a 93% recognition accuracy. As for voicing detection, we used the duration of the unvoiced portion (DUP) of the fricative as the main cue in detecting voicing. Using two physical quantities to extract this feature, an accuracy of 95% was obtained.

The obtained results show significant improvement compared to previous work. Similar experiments, which study the *acoustic-phonetic* automatic recognition of fricatives using multispeaker continuous speech with comparable database size, are quite rare in the literature. Results obtained by Hughes and Halle [12] for place of articulation detection of smaller database size (190 fricatives from 5 speakers) gave between 77%-80% recognition accuracy. They relied mainly on the spectral shape to perform their recognition. The 93% obtained in our experiments indicate a significant improvement that is mainly due to the use of new features, extraction and manipulation algorithms which integrate several acoustic properties in the decision making process. In voicing detection, the obtained results show also a clear improvement over the 83% rate obtained by Stevens *et al* [18]. This is in spite of the fact that the database used was larger, more variable (22 speakers from 5 different dialects versus 3 speakers in their case) and with continuous speech as opposed to the controlled utterances that were used in their experiments. The reason behind this improvement could be attributed to the improved technique used here for detecting phonation (periodicity). This is a clear example of how the translation process from abstract features to physical features could play a significant role in the recognition performance.

The system developed in this work is meant mainly to test and evaluate the features extracted and the algorithms used in their manipulation. Its performance represents a worst-case assessment. When integrated with a system which involves more complicated techniques like training, speaker normalization and variable thresholds, it is expected to give an even better performance.



Is MNSS <threshold (0.02 or 0.01)? OR MDP < 1KHz?
No
Yes
Sibilant
Non-Sibilant (/f/, /v/, /th/ and /dh/)
unvoiced
voiced
/f/ and /th/
/v/ and /dh/
Is the SCG < threshold (4KHz)?
Yes
No
<3KHz?
Location of the largest GSD peak?
>4KHz?
> 3KHz and <4KHz?
<2.5KHz?
Location of the largest GSD slope?
>3.5KHz?
> 2.5KHz and <3.5KHz?
Difference between the largest GSD peak value and the highest order filter GSD output (>3)?
Yes
No
Palatal (/sh/ and /zh/)
Alveolar (/s/ and /z/)
unvoiced
voiced
voiced
unvoiced
/sh/
/zh/
/z/
/s/

Fig.(2). Algorithm for place of articulation detection.

4. ACKNOWLEDGMENT

This work is supported by a grant from Catalyst Foundation.

5. REFERENCES

[1] Ali, A.M.A., “Acoustic Features for the Automatic Recognition of Fricatives”, Technical Report, TR-CST27AUG97, Center for Sensor Technologies, University of Pennsylvania, 1997.

[2] Baum, S. R. and Blumstein, S. E., “Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English”, *J. Acoust. Soc. Am*., 82 (3), pp. 1073-1077, 1987.

[3] Behrens, S. and Blumstein, S. E., “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, pp. 861-867, 1988.

[4] Behrens, S. J. and Blumstein, S. E., “Acoustic characteristics of English voiceless fricatives: a descriptive analysis”, *J. Phonetics*, 16, pp. 295-298, 1988.

[5] Cohen, J.R., “Application of an Auditory Model to Speech Recognition”, *J. Acoust. Soc. Am*., 85, 2623-2629, 1989.

[6] Cole, R. A. and Cooper, W. E., “Perception of voicing in English affricates and fricatives”, *J. Acoust. Soc. Am.*, 58, pp. 1280-1287, 1975.

[7] Cole, R., *et al*, “The Challenge of Spoken Language Systems: Research Directions for the Nineties”, *IEEE Trans. Speech and Audio Proc.*, 3, pp. 1-20, 1995.

[8] Guerlekian, J.A., “Recognition of the Spanish fricatives /s/ and /f/”, *J. Acoust. Soc. Am*., 70, pp. 1624-1627, 1981.

[9] Harris, K. S., “Cues for the discrimination of American English fricatives in spoken syllables”, *Lang. Speech*, 1, pp. 1-17, 1958.

[10] Hedrick, M. S. and Ohde, R. N., “Effect of relative amplitude of frication on perception of place of articulation”, *J. Acoust. Soc. Am*., 94, pp 2005-2026, 1993.

[11] Heinz, J. M. and Stevens, K. N., “On the Properties of Voiceless Fricative Consonants”, *J. Acoust. Soc. Am*., 33, pp. 589-596, 1961.

[12] Hughes, G. W. and Halle, M., “Spectral Properties of Fricative Consonants”, *J. Acoust. Soc. Am*., 28, pp. 303-310, 1956.

[13] Jongman, A., “Duration of frication noise required for identification of English fricatives”, *J. Acoust. Soc. Am*., 85 (4), pp. 1718-1725, 1989.

[14] Klatt, D.H. and Klatt, L.C., “Analysis, synthesis and perception of voice quality variations among female and male talkers”, *J. Acoust. Soc. Am*., 87, pp. 820-857, 1990.

[15] Manrique, A.M.B., and Massone, M.I., “Acoustic analysis and perception of Spanish fricative consonants”, *J. Acoust. Soc. Am*., 69, pp. 1145-1153, 1981.

[16] McCasland, G.P., “Noise intensity and spectrum cues for spoken fricatives”, *J. Acoust. Soc. Am*., Suppl.1, 65, S78-S79, 1979.

[17] Seneff, S., “A Joint Synchrony/Mean Rate Model of Auditory Speech Processing”, *J. Phonetics*, 16, pp. 55-76, 1988.

[18] Stevens, K.N., et al, “Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters”, *J. Acoust. Soc. Am*., 91, pp. 2979-3000, 1992.