

Classification of Voiceless Fricatives through Spectral Moments

Henry Fu

Department of Computer Science
North Carolina State University
Raleigh, NC 27695, USA

And

Robert D. Rodman

David F. McAllister

Donald L. Bitzer

Bowei Xu

Voice Input/Output Group
Multimedia Laboratory
Department of Computer Science
North Carolina State University
Raleigh, NC 27695, USA

ABSTRACT

An algorithm for the classification of voiceless fricatives is presented. Using moments computed from the spectra of speech signals containing voiceless fricatives, the five voiceless fricatives in English—/f/, /th/, /s/, /sh/, and /h/, can be partially separated. Data from three native English speakers are used to develop the algorithm.

Keywords: Voiceless Fricative, Acoustic, Speech and Signal Processing, Phoneme.

1. INTRODUCTION

The purpose of this study is to determine a way to identify the voiceless fricatives of English with just the speech signal. This classification problem stemmed from the study of lip synchronization for facial animation of spontaneous speech.

A voiceless fricative is a speech sound that is made without the vibration of the vocal cords. It is the result of articulators within the mouth cavity coming close together, causing the air to be forced through the narrowed vocal tract. As opposed to a stop, during which the airflow is obstructed for a brief moment before its release, the airflow during the articulation of a fricative is never entirely blocked. This allows fricatives to be separated effectively from stops and affricates [1]. In English, /f/, /th/, /s/, /sh/, and /h/ are the voiceless fricatives. (The initial sounds of "fin," "thin," "sin," "shin," and "him.")

The need to classify voiceless fricatives is that though all fricatives have a high jaw position, each of the aforementioned voiceless fricatives has a noticeably distinct mouth position—with /f/ the lower lip is tucked under the upper teeth, with /th/ the tongue is between the upper and lower teeth, with /s/ the mouth is in a neutral position, with /sh/ the lips are extruded. Before the sounds are processed by the algorithm described below, the voiceless fricatives are extracted manually from their VFV contexts and placed in separate wave files, such that one file contains one fricative in one context. The extraction

and with /h/ the lips anticipate the following vowel. Because of that, accurate classification of the voiceless fricatives is necessary for accurate lip synchronization.

In a previous study [2], our methods of phoneme detection have been successful in separating vowels, voiced fricatives, and (with moderate accuracy) nasals. Though those methods were designed to work with voiced utterances, we were able to apply some of those techniques to achieve a partial separation of the voiceless fricatives.

This paper is organized as follows: section 2 describes the method of data collection; section 3 presents the algorithm that, when applied to the voiceless fricatives, separates them into four categories; the results of this study is presented and discussed in section 4; conclusions drawn from this study are presented in section 5.

2. METHOD

In this study, the voices of three native English speakers (2 males and 1 female) are used. The utterances are recorded under normal room conditions using a microphone connected to a Macintosh computer. SoundEdit 16 version 2 is used to record the utterances. The sound files are saved in WAVE format at 22.050 kHz sampling rate and 8-bit quantization.

The voiceless fricatives are recorded in vowel-fricative-vowel (VFV) contexts. The vowels used are /i/, /u/, and /a/. For example, the /f/ sound is recorded in the following 9 contexts:

[(i) f (i)] [(i) f (u)] [(i) f (a)]
[(u) f (i)] [(u) f (u)] [(u) f (a)]
[(a) f (i)] [(a) f (u)] [(a) f (a)]

criteria are based on amplitude, zero-crossing rate, and periodicity of the sound signal in the time domain. This is because voiceless fricatives are not periodic, generally have high zero-crossing rates, and have amplitude ranges

significantly different from that of the surrounding vowel sounds. The sampling rate and quantization remain unchanged for the resultant wave files (45 per speaker).

3. ALGORITHM

Overview

The voiceless fricative classification algorithm involves analyzing the shape of the spectrum produced by computing the discrete Fourier transforms on 100-sample windows. The measure of shape is derived by treating the spectrum as a probability density function and computing statistical moments. The values of two of the moments (the first moment and the second central moment) are used to classify the fricatives.

The Steps

The following is a nine-step process for the classification:

1. compute the discrete Fourier transform (DFT) on 100 samples of the signal to transform it from the time domain to the frequency domain
2. take the absolute value of the result and scale by dividing it by the square root of the window width, which is 10 in a 100-sample window
3. shift over 1 sample
4. do steps 1 through 3 one hundred times
5. average those 100 transforms, scale again by taking the cube root of that average to reduce the effect of the first formant, drop the DC term, and interpolate it with a degree 3 polynomial curve to produce a spectrum
6. convert the spectrum to a probability density function by dividing it by its mass, then calculate the first moment (mean) and the second central moment (variance) of that function in the range of 0 to 8000 Hz, and put them in two lists
7. repeat steps 1 through 6 until less than 300 samples remain
8. scale and smooth the two lists of moments
9. the two lists are then each averaged, producing a representative point for the voiceless fricative in the vowel-fricative-vowel (VFV) context

Explanation

Starting from the beginning of the signal, take 100 samples at a time and do as follows:

take sample 1 – 100, compute its DFT
take sample 2 – 101, compute its DFT
take sample 3 – 102, compute its DFT
...
take sample 100 – 199, compute its DFT

For each DFT, the absolute value is taken and the result divided by the square root of the window width, as described in step 2 above. The square root of the window width is always 10 since a fixed window of 100 samples is used. Average those 100 transforms and take the cube root. This has the effect of lessening the influence of the first formant and accentuating the differences that are relevant in other parts of the spectrum. The DC term is then dropped and the discrete points are connected by a third-degree polynomial interpolation. The now continuous DFT is converted to a probability density function by dividing by its mass. The mean and the variance of the first 8000 Hz of the spectrum is

calculated and stored in two separate lists. The reason for the cut-off at 8000 Hz is that the spectra below that value provide adequate separation of the voiceless fricatives. This process is then repeated, starting with sample number 101, and so on, until less than 300 samples remain in the input signal. The last few hundred samples of the utterance are not processed because of the increasing vowel influence on the fricative.

The two lists of moments are then scaled (the mean is divided by 10^3 , and the variance is divided by 10^6), smoothed (average of 9), and averaged, producing a point in moment space (mean vs. variance). That is the “representative” point in moment space for the voiceless fricative in the particular context. Voiceless fricatives in other contexts are processed the same way.

4. RESULTS

By processing the voiceless fricatives in the above manner, they separate into four distinct regions in the mean-variance plot that is somewhat speaker independent (refer to figures 1 through 3, and note that θ is the IPA symbol for /th/ and \int is the IPA symbol for /sh/):

- The f/th-region, consisting of the voiceless fricatives /f/ and /th/.
- The s-region, consisting of the voiceless fricative /s/.
- The sh-region, consisting of the voiceless fricative /sh/.
- The h-region, consisting of the voiceless fricative /h/.

For all three speakers, the s-region is always to the right of the other regions, so the /s/ sounds can be separated from the others by the mean (first moment) of the spectrum. That threshold is approximately 3.9 to 4.3, and varies slightly for different speakers.

With the /s/ sounds identified and separated from the other voiceless fricatives, the region for the /sh/ sounds is always below the region consisting of /f/ and /th/ sounds for all three speakers. That is, the second central moment (variance) is smaller for /sh/ than for /f/ and /th/. The threshold for Speaker R (figure 1) is between 4.3 and 4.95, for Speaker D (figure 2) is between 5.15 and 5.35, and for Speaker N (figure 3) is between 5 and 5.35. So it varies a bit for different speakers. A clue that is useful in distinguishing /sh/ sounds from /f/ or /th/ sounds for a speaker is by looking at the location of the s-region for that speaker. The s-region is closer to the sh-region for all three speakers.

The region for the /h/ sounds is to the left of the f/th-region and the sh-region, and on the vertical (variance) axis, /h/ sounds are situated between f/th- and sh-regions (figures 1 through 3). Applying a clustering algorithm with the /h/ centroid to the left and above the sh-region will capture most of the representative points for /h/ sounds. Since f/th-, s-, and sh-regions are well separated in most cases, the same clustering algorithm can also be applied to those three regions.

The results presented here are consistent with the observations made in [3], where it states that /f/, /th/ and /h/ each has a low-energy, diffuse spectrum, resulting in the large values in the second moment (variance) of the spectrum; /s/ tends to have a high mean; and /sh/ has mean that is lower than that of /s/.

In this study, it is also found that by comparing moments of wider spectrums, i. e. 0 to 11000 Hz, there is a general trend of tighter clustering of the same voiceless fricative from the same speaker, while the regions for different voiceless fricatives move farther apart. Those results are shown in figures 4 through 6.

5. CONCLUSIONS

While most voiceless fricatives can be separated by comparing their spectral moments, the sounds /f/ and /th/ still present a problem. What remains now is to devise a procedure to distinguish /f/ and /th/. It is crucial in the context of lip sync, since those two voiceless fricatives have significantly different lip positions.

To address the /f/-/th/ problem, we are currently carrying out analysis on the entire sound (all of VFV, not just the voiceless fricative), which allows the processing of the transitions from vowel to fricative and from fricative to vowel. That is because other studies have indicated that the difference between /f/ and /th/ is at the transitions into and out of the voiceless fricative [4].

6. REFERENCES

- [1] L. F. Weigelt, S. J. Sadoff, and J. D. Miller, "Plosive/Fricative Distinction: The Voiceless Case", *Journal of the Acoustical Society of America*, Vol. 87, No. 6, June, 1990, pp. 2729-2737.
- [2] D. F. McAllister, R. D. Rodman, D. L. Bitzer, A. W. Freeman, "Speaker Independence In Automated Lip-Sync For Audio-Video Communication", *Computer Networks and ISDN Systems*, Vol. 30, Nos. 20-21, 1998, pp. 1975-1980.
- [3] N. J. Lass, "Principles of Experimental Phonetics", St. Louis, Mo.: Mosby, 1996.
- [4] R. Smits, "Human Consonant Recognition For Initial And Final Segments of VCV Utterances", URL: <http://pitch.phon.ucl.ac.uk/home/sh110/roel/real.htm>.
Department of Phonetics and Linguistics, University College London.

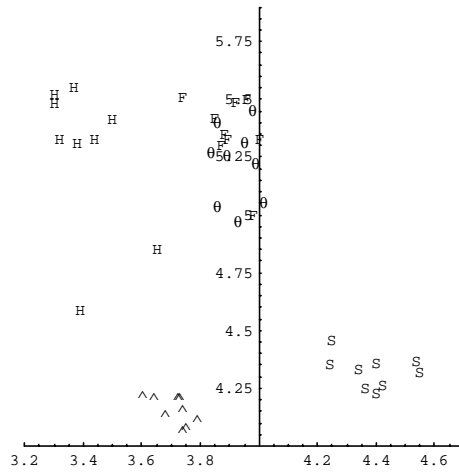


Figure 1: Voiceless Fricative Spectral Moments (0 to 8 kHz) for **Speaker R**; (note that θ is the IPA symbol for /tʰ/ and \int represents /ʃ/)

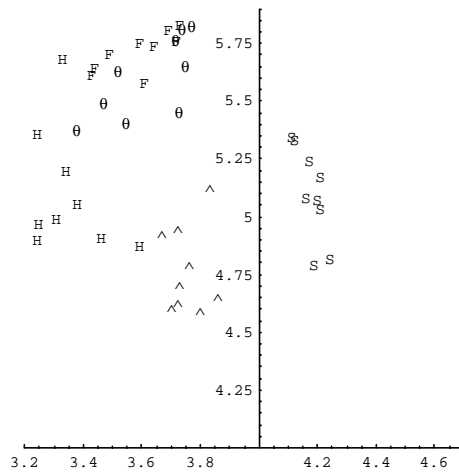


Figure 2: Voiceless Fricative Spectral Moments (0 to 8 kHz) for **Speaker D**

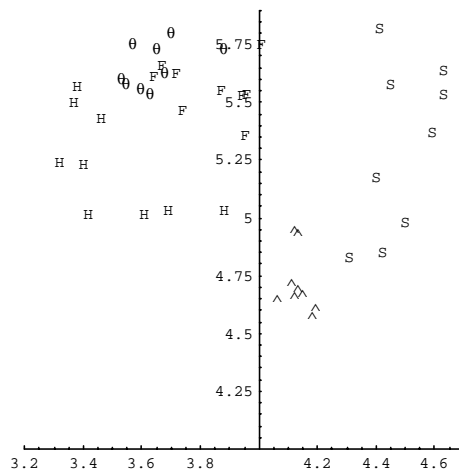


Figure 3: Voiceless Fricative Spectral Moments (0 to 8 kHz) for **Speaker N**

Voiceless Fricative Spectral Moments, 0 to 8 kHz

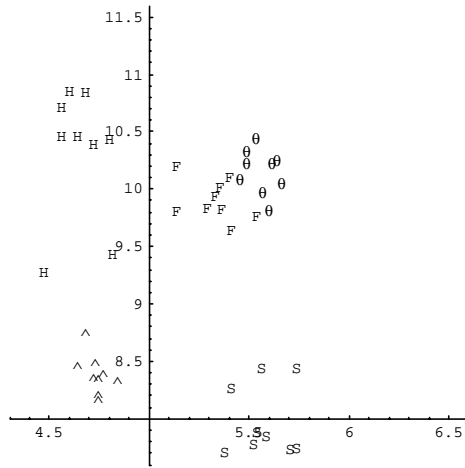


Figure 4: Voiceless Fricative Spectral Moments (0 to 11 kHz) for Speaker R

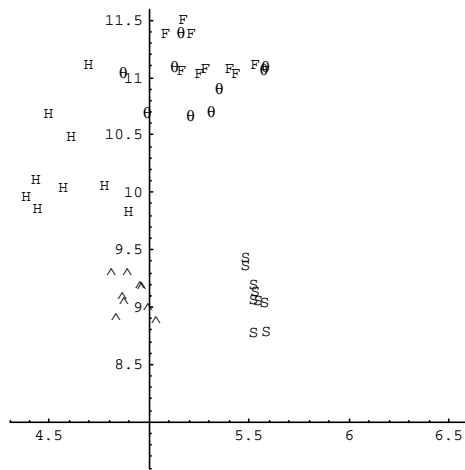


Figure 5: Voiceless Fricative Spectral Moments (0 to 11 kHz) for Speaker D

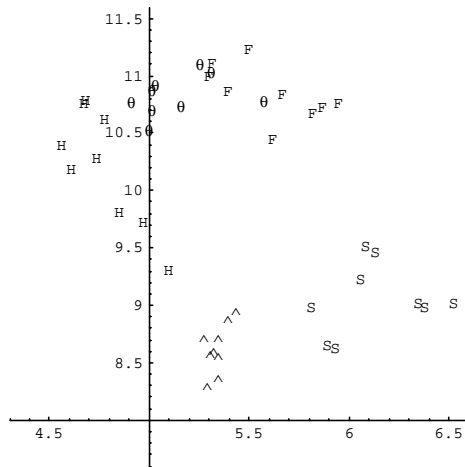


Figure 6: Voiceless Fricative Spectral Moments (0 to 11 kHz) for Speaker N

Voiceless Fricative Spectral Moments, 0 to 11 kHz