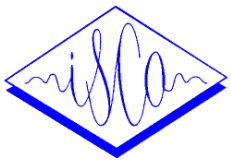

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Distinguishing the voiceless fricatives F and TH in English: a study of relevant acoustic properties

Kazue Hata, Heather Moran and Steve Pearson
Speech Technology Laboratory, Panasonic Technologies, Inc.
3888 State Street, Santa Barbara, California 93105, USA

ABSTRACT

Distinguishing between the voiceless fricatives F and TH is a difficult problem for both natural speech and synthetic speech. We report the results of experiments and spectral analyses designed to find distinguishing acoustic characteristics of the voiceless fricatives F and TH. These experiments were also designed in consideration of our hybrid text-to-speech system, which combines formant synthesis with concatenated units from natural speech. In our system, the intelligibility of initial stops has improved dramatically in comparison with the formant-synthesizer-only version of our system, but F and TH are still highly confusable [1]. In this study, we used only natural speech, and conducted perceptual experiments by using frication-only stimuli and stimuli combining frication with segments of the following vowel.

The results showed that when a frication portion alone was presented, F was correctly identified more often than TH. When the frication portion with the entire following vowel was presented, the identification of F showed significant improvement, reaching more than 90% for the /a u/ vowel contexts, and for the /i/ context, increasing from 40 to 78%. By contrast, the identification of TH failed to show any significant improvement.

I. INTRODUCTION

Among fricatives, F and TH, the noise sources of which are the least Intense, have been reported to be highly confused [2,3]. In our new hybrid text-to-speech system, F and TH proved to be the most confusable. TH was perceived as F in word-initial and -final positions more often than F was perceived as TH [1]. For this study we conducted perceptual experiments and examined spectral shapes and vowel transitions of these two fricatives with hopes of finding distinguishing acoustic characteristics. Ultimately, we plan to apply the knowledge gained to our synthesis system to improve their intelligibility.

Many researchers have looked for distinctive cues in the noise source itself, and/or adjacent vowels. Hughes and Halle [4] reported that F was characterized by a considerable low frequency energy peak less than 2 kHz and sometimes by a peak higher than 10 kHz. Strevens [5] and Abbs and Minifie [6] indicated that the spectral peaks of F and TH are located around 3 kHz. Heinz and Stevens [7] found that the addition of synthetic vowels to the frication portions still resulted in F-TH confusions. Although no specific data were provided, they indicated that F2 transition seemed to be a primary cue for F-TH distinction.

Harris [8] was the first researcher to demonstrate that unlike S and SH, vowel transition is a primary cue for the perceptual distinction between F and TH. She concluded that the listeners depend on the F2 transition of the adjacent vowels, which is related to tongue forward-backward movement. Similar results were obtained by McCasland [9], who conducted a splicing experiment with the vowel context /i/. He found that these fricatives can be distinguished from each other by the formant transition cues of the vocalic portion of the syllables.

Other experiments have explored the role of amplitude. Heinz and Stevens [7], by synthesizing fricatives, found that F and TH required very high frequency peaks on the order of 8 kHz, and for these fricatives, the frication noise needed to be at least 10 dB lower than the vowel amplitude, with 25 dB being optimal. Guerlekian [10] showed that low noise amplitude relative to the vowel was perceived as F, and high amplitude noise, as S. McCasland [11] also found this relative noise amplitude to be the primary characteristic distinguishing F and TH from S and SH. The same tendency was found by Stevens [12] who varied the amplitude of F5 and F6 for S and TH. He found that when the high-frequency noise amplitude decreased relative to the vowel, TH was perceived more, whereas when increased relative to the vowel, S was perceived more.

Shadle et al. [13] examined both noise spectra and vowel formant transitions with fricative-vowel-fricative sequences. Using an ensemble averaging technique, they found that vowel context had a clear effect on fricative spectra. By dividing the fricatives into three temporal phases, they found that while the mid steady-state phase of the frications was similar, distinctive cues for F and TH were located in the vowel-fricative and fricative-vowel transition phases, where TH has a higher amplitude than F. Further, F2 and F3 for TH increase relative to those of F, supporting the claim that formant transitions are the distinctive cues to these fricatives.

The purpose of this study was to find acoustic characteristics distinguishing F and TH, which can be implemented to improve our hybrid text-to-speech system. We conducted two perceptual experiments. In the first experiment, we investigated whether spectral cues alone are sufficient to distinguish F and TH in some vowel contexts. We followed a procedure of one of the experiments on strong fricatives (S SH Z ZH) conducted by Yeni-Komshian and Soli [14], who presented these frication stimuli in three different vowel contexts (/i a u/) to subjects. In the second experiment, we compared frication-only stimuli with stimuli containing a vowel segment or the entire vowel to investigate whether the adjacent vowels provide perceptual cues for F and TH.

II. FIRST EXPERIMENT

1. Procedure

This experiment investigated the perception of spectral variations in the fricatives F and TH produced in the different vowel contexts /i a u/. A female native speaker of American English read a list of elongated fricatives in isolation and a list of CVC words with F, TH, S, SH or H in word-initial position, and F, TH, S, or SH in final position (since no English word ends with H, this fricative was omitted). F and TH were the main focus but the other fricatives, presented only in the /i a/ contexts, were included for comparison.

The fricative portion without any periodic vowel waveforms was extracted from these CVC sequences. The signal was downsampled from the original DAT recording to 10 kHz, the sample rate of our synthesizer. There were three types of fricative stimuli: one extracted from the initial fricative, another from the final fricative and the one uttered in isolation. The initial fricative stimuli were 120 msec in duration excised back from the point of voice onset of the following vowel. Since the duration of the final fricative tended to be longer, the final fricative stimuli were 150 msec excised from the point of voicing offset. The duration of the stimuli uttered in isolation was 150 msec excised from the middle of elongated fricatives. No amplitude normalization was conducted to compensate for variations in the frication amplitude.

For the listening task, we used five repetition of the F an TH stimuli and three repetitions of the other fricative stimuli, and the resulting 125 stimuli were randomized and presented to twenty native speakers of American English (six males and fourteen females). Part I of this experiment was intended to test word-initial fricatives and Part II, word-final fricatives. The stimuli uttered in isolation were included in both parts. The subjects were asked to listen to each sound through headphones and to choose the one perceived from the five fricatives in Part I and the four in Part II on the answer sheets.

2. Results and Discussion

Tables 1-a and 1-b show the percentages of identifications for F and TH stimuli, respectively. The rows show the type of stimuli being presented, while the columns show subjects' responses to those stimuli.

Table 1-a: Identification (%) for initial F (Part I)

	F	TH	S	SH	H
F(i)	20	42	27	11	0
F(a)	44	18	30	4	4
F(u)	47	21	28	4	0
isolated F	54	21	20	3	2

Table 1-b: Identification (%) for final F (Part II)

	F	TH	S	SH
(i)F	42	34	21	3
(a)F	42	32	21	5
(u)F	56	29	9	6
isolated F	47	21	32	0

Tables 1-a and 1-b show that F stimuli were identified as F more frequently than as any of the other fricatives except for one case. In the case of the initial F followed by /i/, it was identified as TH 42% of the time. The identification scores for F were rather low in comparison with S (55% for initial and 88% for final positions) and SH (88% for initial and 94% for final).

In addition, F was identified as TH and S at more than a chance level. (The chance probability of occurrence for each entry in Tables 1-a and 1-b is 20% and 25%, respectively.) Initial F followed by /a/ or /u/ and in isolation was often perceived as S, whereas final F in the three vocal contexts was perceived as TH. The difference between these first and second candidates for F were statistically significant ($p < .05$) for only two contexts: F in isolation in Part I and F preceded by /u/ in Part II. Thus, F was perceived as F often, but we can conclude that its identification is not as certain to the listener as strong fricatives, at least when limiting the spectra to 0-5 kHz.

Table 2-a: Identification (%) for initial TH (Part I)

	F	TH	S	SH	H
TH(i)	42	22	33	3	0
TH(a)	48	24	24	3	1
TH(u)	32	40	23	5	0
isolated TH	65	16	14	5	0

Table 2-b: Identification (%) for final TH (Part II)

	F	TH	S	SH
(i)TH	39	17	37	7
(a)TH	33	34	30	3
(u)TH	53	21	22	4
isolated TH	71	9	14	6

Tables 2-a and 2-b show that the identification of TH was very different from that of F: the TH stimuli were more often identified as F than TH. Also this identification as F fluctuated between 30% and 70%. Interestingly, isolated TH was perceived as F more than 65% of the time. In addition to the misperception of TH as F, TH was perceived as S in the /i/ contexts above a chance level (33% for initial TH and 37% for final TH).

In order to understand the perceptual confusions of F and TH, a spectral analysis was conducted for the original DAT recording. Figures 1 and 2 show the spectra shapes of fricatives downsampled to 16 kHz. We examined the spectra of 30-msec of frication starting 150 msec into elongated isolated fricatives. For initial and final fricatives, the 30-msec of frications adjacent to the vowel was used to create the spectra.

Only minor differences between F and TH can be found in Figure 1. The energy peaks of the TH in this figure shows the same amplitude values as those of the F spectra. In Figure 2, the peaks in TH were 5-10 dB lower than those in F across different vowels. This possibly implies that when the speaker uttered these fricatives in isolation, she made some loudness adjustment for TH. In addition, in Figure 1, the first energy peaks for both F and TH occurred near 2 kHz and the second ones near 4 kHz. The third peak, which occurred near 6 kHz, was very prominent for TH, whereas the energy above 6 kHz was comparatively flat. Since the stimuli used in the experiment carried no information above 5 kHz, the spectral similarities in the 0 to 5 kHz region demonstrated in Figure 1

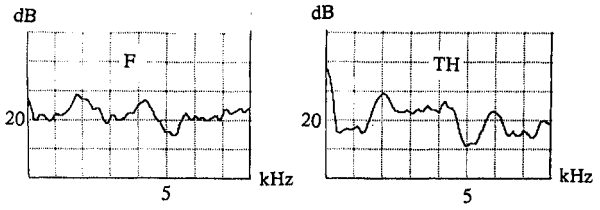

Figure 1: Spectra shapes of F and TH uttered in isolation

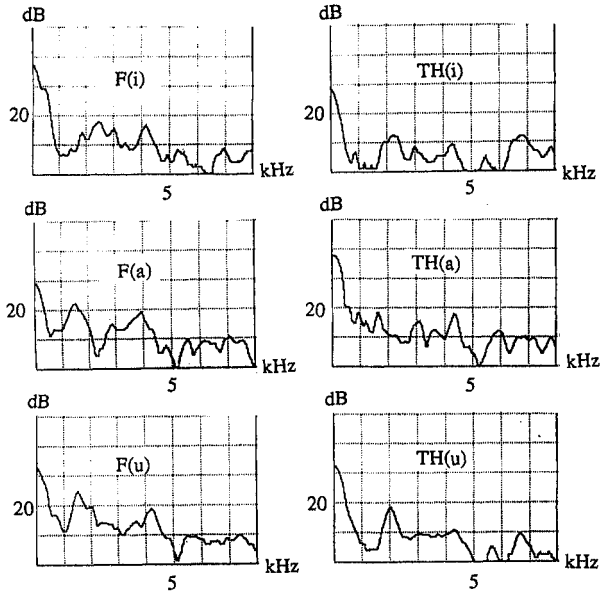

Figure 2: Spectra shapes of initial F and TH in three vowel contexts

may have triggered perceptual confusions between F and TH uttered in isolation.

The general trend that both F and TH failed to obtain as high identification scores as other fricatives such as SH could be explained by the lack of the relative amplitude cue [7, 10, 11, 12]. There was an overall amplitude difference between F and TH seen in Figure 2. Hendrick and Ohde [15] claimed that the frication energy peak is a primary cue to place of articulation of fricatives while the comparison between the frication peaks and a vowel peak in the same frequency region has secondary perceptual importance. In the first experiment, since the frication was presented without a following vowel, we suspected that the subjects could not use, as a cue, the amplitude of frication noise relative to vowel onset. Thus, it was necessary to test the importance of the following vowel in the perception of these fricatives.

As far as the confusion of TH as F is concerned, we were unable to find any significant generalization across vowel contexts, though we did find the vowel effect on the fricative spectra in Figure 2. These findings in the first experiment led us to run another experiment.

III. SECOND EXPERIMENT

1. Procedure

The second experiment was prepared to study the perception of F and TH, using only word-initial fricatives with the same three vowel contexts as the first experiment. This time, we downsampled the original DAT recording to 16 kHz, to include the spectral information in the higher spectral region which contained cues for F and TH distinction as shown by our analysis. The stimuli were excised in the frication noise 120 msec back from the voice onset, as in the first experiment. For the offset of the frication noise, we made three types by excising at: (1)the voice onset, (2) 30 msec into the voice onset and (3) the end of the vowel. Eighteen combinations (2 fricatives x 3 vowels x 3 offset types) were repeated four times and randomized. The 72 stimuli were presented to ten subjects, who were asked to make a forced choice between F and TH.

2. Results and Discussion

As seen in Tables 3 and 4, there was a difference between F and TH in the effect of a vowel segment and the entire vowel on the fricative identification. The identification scores of F followed by the entire vowel were significantly different ($p < .05$) from the ones of frication-only stimuli. More than 90% accuracy was achieved in the /a/ and /u/ context. However, the addition of a 30-msec vowel transition was not statistically significant in any vowel context due to a wide variation found in the subjects' responses, though the scores increased more than 15% in comparison with frication-only stimuli.

By contrast, for TH, the addition of a vowel segment decreased or failed to improve the identification. For instance, When followed by /a/, the frication-only stimuli reached a 60% correct identification. The addition of a 30-msec vowel decreased the score to 35%. The differences in scores in these two conditions were not statistically significant.

Table 3: Correct identification (%) for F according to different vowel durations

	F(i)	F(a)	F(u)
no vowel	40	65	53
30-msec vowel	55	85	68
entire vowel	78	93	100

Table 4: Correct identification (%) for TH according to different vowel durations

	TH(i)	TH(a)	TH(u)
no vowel	75	60	80
30-msec vowel	75	35	70
entire vowel	58	58	78

In the experiment, the vowel contexts did not seem to play an important role in the identification of TH, unlike the suggestion by many researchers such as Harris [8]. In a preliminary study, we conducted a similar experiment with 10 kHz sample rate. The identification of TH increased with 16 kHz sampling in six stimuli types by 5-25%. However, these

differences between 10 kHz and 16 kHz were not statistically significant. Thus, our finding in TH confirmed the results obtained by Jongman [16] who examined the fricatives from the viewpoint of frication duration. He showed that 50-msec duration of frication from the fricative onset is sufficient for most fricatives to be identified. The only exception is that as for TH reasonably accurate perception was achieved when presented in full frication or with the following vowel. However, even when the vowel transitions and steady-state vowel were added, its identification was not significantly different from the frication-only stimuli.

IV. SUMMARY

The distinction between the fricatives F and TH is a difficult problem with either natural or synthetic stimuli. Unfortunately, F and TH have not been a subject of a wide range of phonetic and perceptual research, so more research needs to be performed before we can apply such results to improving the intelligibility of our text-to-speech system.

In this study we conducted experiments on fricative identification considering three vowel contexts and two different sampling rates. We found that when the frication portions of the stimuli were presented in isolation, F was correctly identified more often than TH.

For the identification of F, the addition of the entire vowel increased its identification from less than 66% to more than 90% in the /a/ and /u/ contexts, while the /i/ context the identification rate almost doubled, from 40% to 78%. Using just the initial 30-msec of the vowel segment, however, failed to provide a significant cue to this fricative. We thus found that to obtain salient perceptual cues from the following vowel more than 30 msec of the vowel is required, and possibly the entire vowel is required. In order to determine the vowel duration required to obtain good cues, we are planning to run another experiment with stimuli with more steps of vowel duration. We plan to implement these results into our hybrid text-to-speech system by more accurately modeling the fricative/vowel coarticulation in different contexts, and for the required duration.

The identification of TH remains elusive with both 10 kHz and 16 kHz sampling. Surprisingly, adding either the entire vowel or the initial 30-msec portion failed to improve identification in either of the three vowel contexts. Perhaps we obtained these results because the stimuli in the experiments did not contain enough of the beginning phase of a TH fricative, so the fricative onset may have sounded unnatural. In the future, we plan to thoroughly investigate this beginning phase of the fricatives that was not well-represented in the current study. Perhaps the combination of different phases such as the initial, mid steady-state, and final release phases, as well as transition phases, are important for the identification of the weak fricatives.

Finally, in our experiments, we limited the spectral bandwidth of the stimuli to the 0-5kHz and the 0-8 kHz region. In the future we plan to consider the effects of a much higher frequency for these weak fricatives, as suggested in some studies [4,7,13].

V. REFERENCES

[1] Moran, H., K. Hata, and S. Pearson. 1994. The use of sampled consonants for improved intelligibility in formant synthesizers. JASA Pt. 2. 2816.

[2] Miller, G.A. and P.E. Nicely. 1955. An analysis of perceptual confusions among some English consonants. JASA 27.338-352.

[3] Kuehn, D.P. and K.L. Moll. 1972. Perceptual effects of forward coarticulation. J. of Sp. Hear. Res. 15, 654-664.

[4] Hughes, G.W. and M. Halle. 1956. Spectral Properties of Fricative Consonants. JASA 28.303-310.

[5] Strevens, P. 1960. Spectra of fricative noise in human speech. Language and Speech, 3.32-49.

[6] Abbs, M.S., and F.D. Minifie. 1969. Effect of acoustic cues in fricatives on perceptual confusions in preschool children. JASA, 46.1535-1542.

[7] Heinz, J.M., and K.N. Stevens. 1961. On the properties of voiceless fricative consonants. JASA 33.589-596.

[8] Harris, K.S. 1954. Cues for the identification of the fricatives of American English. JASA 26.952.

[9] McCasland, G.P. 1979a. Noise intensity and spectrum cues for spoken fricatives. JASA Suppl. 1 65, S78-S79.

[10] Guerlekian, J.A. 1981. Recognition of the Spanish fricatives /s/ and /f/. JASA 70.1624-1627.

[11] McCasland, G.P. 1979b. Noise intensity cues of spoken fricatives. JASA Suppl. 1 65, S88.

[12] Stevens, K.N. 1985. Evidence for the role of acoustic boundaries in the perception of speech sounds. In: V.A. Fromkin (ed.), Phonetic Linguistics: Essays in Honor of Peter Ladefoged. New York: Academic Press. pp. 243-255.

[13] Shadle, C.H., A. Moulinier, C. U. Dobelke and C. Scully. 1992. Ensemble averaging applied to the analysis of frication consonants. Proc of ICSLP, 1992, Kobe, Japan. pp.53-56.

[14] Yeni-Komshian, G., and S.D. Soli. 1981. Recognition of vowels from information in fricatives: perceptual evidence of fricative-vowel coarticulation. JASA 70.966-975.

[15] Hendrick, M.S. and R.N. Ohde. 1993. Effect of relative amplitude of frication on perception of place of articulation. JASA 94.2005-2026.

[16] Jongman, A. 1989. Duration of frication noise required for identification of English fricatives. JASA 85.1718-1725.