ACOUSTIC PROPERTIES AT FRICATIVE-VOWEL BOUNDARIES IN AMERICAN ENGLISH

L. F. Wilde and C. B. Huang[1]

Research Laboratory of Electronics and Department of Electrical Engineering and Computer Science,
Massachusetts Institute of Technology, Cambridge, MA, USA

ABSTRACT

Previous work has shown that acoustic properties which signal place of articulation and voicing for fricative consonants can be located at the fricative-vowel boundary. Therefore, we focused on the region within 30 ms of the boundary in our search for acoustic regularities.

Our goal is to better understand the perceptual salience of these acoustic cues. In this paper, results will be presented from perceptual tests with natural and synthetic CVs. As expected, the non-strident fricatives were most often confused and the performance for synthesized fricatives was poorer than that for natural speech. Acoustic evidence for these results is examined.

1. INTRODUCTION

Previous acoustic analysis of natural speech has shown that acoustic properties in the vicinity of fricative-vowel boundaries can be associated with cues for consonant perception. The following acoustic attributes are associated with fricative production: (1) an interval of frication noise with a spectrum that is shaped by the location of the constriction in the vocal tract; (2) formant transitions into adjacent vowels that provide additional place of articulation information; and (3) details in the transition from noise production to voicing onset which signal the distinction between voiced and voiceless fricatives. [4].

Our goal is to better understand the acoustic properties of the fricative-vowel boundary and, particularly, their perceptual relevance for place. Describing fricative-vowel boundaries is an especially interesting problem because these occur between continuous sounds produced by different source mechanisms: the supraglottal source, frication noise, which is generated as air flows through a narrow contstriction in the vocal tract, and the two glottal sources, voicing and aspiration.

Fricative synthesis provides a means for systematically examining the relative timing of the different sound sources. By comparing perceptual and acoustic measures for natural and synthetic stimuli, we can evaluate the adequacy of existing rules for modelling fricatives. We performed a series of listening tests to provide a baseline for intelligibility of natural fricatives and fricatives produced by a high-quality speech synthesizer. We also examined the acoustic properties of these stimuli to determine which differences could account for observed deficiencies in intelligibility.

2. PERCEPTUAL TESTS

2.1 Objective

Identification tests were run with natural CV speech tokens as stimuli to provide a baseline measure of intelligibility of fricatives. We used one of the best existing rule-based, text-to-speech synthesizers available to obtain corresponding synthetic stimuli. Nevertheless, we expected that the identification of the synthesized speech would be more difficult.

[1]Presently affiliated with Dragon Systems, Inc., Newton, MA, USA.

2.2 Method

The natural stimuli were CV tokens excised from C'VCVC'VC nonsense utterances spoken by one male speaker, Dennis Klatt [4]. The bandwidth of these utterances, which were digitized at a 10 kHz sampling rate, corresponds to the bandwidth used for synthesis of male speech. The C was one of the eight English fricatives, which can be classed according to place: labiodental (/f/, /v/), dental (/th/, /dh/), alveolar (/s/, /z/), and palatal (/sh/, /zh/). The V was one of four American English vowels (/iy/, /eh/, /aa/, /uw/), chosen to be representative of front, back and rounded vowels. The vowels were truncated 40 ms after vowel onset, which was defined as the beginning of the first identifiable pitch pulse for voiceless fricatives and the point where voicing amplitude increases abruptly for voiced fricatives.

Corresponding CV stimuli were synthesized using the phoneme input mode of KLATTALK, a research version of Klatt's text-to-speech system. The KLATTALK algorithm for formant transitions begins by looking at the segment following the fricative. Values from previous trial and error matching of natural frication spectra were used to optimize table values for synthesis of frication [4]. We used parameter values that Klatt chose to model his own voice.

The voiceless (/f/, /th/, /s/, /sh/) and voiced (/v/, /dh/, /z/, /zh/) fricatives were presented in separate identification tests in each of the natural and synthetic conditions. Five repetitions of the 20 distinct stimuli in each test were presented in random order over headphones in a sound-treated room. Five phonetically trained listeners acted as subjects. Subjects were asked to identify the fricatives by making a forced choice among the four possibilities. No responses regarding the vowel identities were required and voiced-voiceless distinctions were not explicitly examined.

2.3 Results

All of the natural and synthetic strident fricative tokens were identified correctly (0 errors out of 800 total responses). The non-strident fricatives were most often confused (175 errors out of 800 responses). Figure 1 compares the percentage of errors made on the natural and synthetic non-strident fricative CVs in the voiceless and voiced tests.

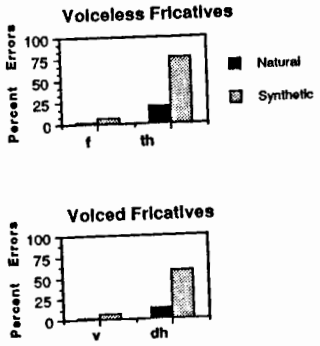

Figure 1. Error distribution, combined across vowel contexts, as a percentage of 100 responses (5 repetitions X 4 vowels X 5 listeners)

As expected, the performance for synthesized fricatives was consistently poorer than natural speech. These results also show that the predominance of errors was for the synthetic /th/ and /dh/ tokens.

3. ACOUSTIC ANALYSIS

3.1 Objective

We examined the acoustic properties of the natural and synthetic stimuli to determine which differences could account for the observed deficiencies in intelligibility. Following Klatt[4], we focused on the following attributes involved in moving from a frictive to a vowel: the evidence of the changing sound sources (voicing, frication, aspiration) and the onset frequency of formants (F1, F2, F3).

3.2 Method

All measurements were performed with the set of tools for speech analysis available on the MIT Speech Vax cluster. The formant onset frequencies were measured at the first identifiable pitch pulse. Discrete Fourier transforms were calculated with a 6.4 ms Hamming window that was carefully placed in order to maximize inclusion of the closed portion of waveform.

The noise spectra were calculated over a 30 ms Hamming window. Spectra were computed at 10 ms intervals from consonant onset to the fricative-vowel boundary. We compared relative amplitudes of noise in different frequency regions to vowel formant amplitudes.

3.3 Results

In view of perceptual test results, the acoustic findings for the non-strident fricatives only are presented here.

A close correspondence was found between natural and synthetic tokens for the first three formant onset values. As previously seen in the high front vowel context, F2 onset formant frequencies contradict the general rule that formant frequency is always lowest for labial place of articulation [4]. Figure 2 illustrates the F2 formant onset frequencies for voiceless natural and synthetic fricatives.

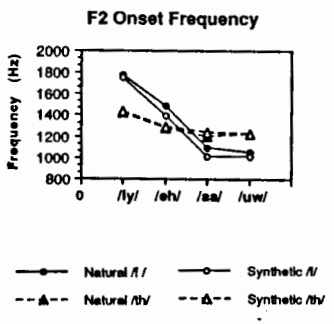

Figure 2. F2 onset frequencies for voiceless fricatives /f/ and /th/ in each vowel context.

Evidence of aspiration, if present, was usually found within 20 ms before the vowel onset. Aspiration may be distinguished from frication by an F2 prominence in the noise spectrum continuous with the formant at the fricative-vowel boundary. Aspiration was indicated, according to this criterion, for all the natural /f/ stimuli. Aspiration was not seen for the natural /th/ stimuli, except for /thuw/. The synthesized stimuli did not include any aspiration in the parameter specifications.

In examining the noise spectra prior to the fricative-vowel boundary, we concentrated on the non-strident token pairs whose difference in intelligibility between the natural and synthesized stimuli was largest (/feh/-/theh/ and /veh/-dheh/). The natural /f/ and /th/ differed most noticeably in their average amplitudes, as compared between the approximate noise amplitude in the 2000-3000 Hz region 70 ms before vowel onset relative to the amplitude of F3 in the vowel (2500 Hz). The natural /f/ was 20 dB lower, whereas /th/ was 30 dB lower than the vowel. The synthetic /f/ was 25 dB weaker while the /th/ was only 20 dB weaker. Time-varying spectral amplitude was observed for the natural /f/, which increased by approximately 10 dB from its beginning to the vowel onset, whereas the spectral amplitude for the natural /th/ appeared constant throughout its duration.

The spectral characteristics of the natural /f/ and /th/ tokens were found to be otherwise similar to each other, consistent with findings in previous work with a larger number of tokens and speakers [2]. The synthesizer models the spectrum of the non-strident fricatives as Gaussian noise with no formant structure. The spectra of the synthesized /f/ and /th/ were tilted, emphasizing the low frequencies with a 6 dB roll-off, whereas the natural stimuli were flat.

The spectra for the natural voiced non-strident fricatives (/v/ and /dh/) were very different from those of the synthesized stimuli. In the natural stimuli, the formant structure extended far into the fricative and noise excitation coexisted with essentially vocalic-looking formant structure. In constrast, there was no region with strong formant structure in the synthetic voiced fricative; instead there was an abrupt change between noise excitation and prevoicing by the glottal source. For /dh/, the onset of the vowel from the prevoiced region was abrupt enough to appear stop-like.

4. DISCUSSION

All of the strident fricatives, which are characterized by a relatively high spectrum amplitude as compared to the adjacent vowel, were identified correctly.

Harris [3] found that the frication noise provides the dominant cue for discriminating /s/ and /sh/, but that formant transition cues dominate the differences in noise spectra for /f/ and /th/.

The need to further investigate distinctions between the labiodental and dental fricatives is highlighted by the current perceptual and acoustic results. The F2 formant frequency onsets in front, back and rounded vowel contexts in the present database illustrate that listeners may adapt to regularities in formant onset frequencies, even if these present unexpected patterns. One possible explanation for the lower F2 formants for /th/ as compared to /f/ before high front vowels is that for labiodentals, the tongue is freer to move in anticipation of the following vowel.

The current findings suggest that even if the formant transitions are reproduced accurately, as in the Klattalk stimuli, deficiencies in intelligibility for synthetic stimuli remain. This implies that further consideration of noise amplitude and shape is needed. While the natural /th/ was 5-10 dB weaker than /f/ relative to the vowel, the synthetic /th/ was too strong and the /f/ too weak to maintain this distinction. This difference could partially explain confusions between synthetic non-strident tokens.

Acoustic variations can be interpreted with respect to existing production models [1] and predictions regarding the interaction and relative amplitudes of frication noise, aspiration, and voicing as constriction sizes vary in time [6]. Our analysis results for the natural and synthetic stimuli suggest the need to better model these source changes between the vowel and the fricative.

In some voiceless fricatives, aspiration can lead to a smoother transition at the fricative-vowel boundary. The role of aspiration in analysis, synthesis and perception of voice are described in Klatt and Klatt[5]. We intend to use the KLSYN88 formant they describe, which provides more flexible control over the glottal source, to continue to model the acoustic characteristics we observed. We can then test if the extra formant structure present in aspiration may provide place-of-articulation information and thus enhance both intelligibility and naturalness.

After synthesizing new tokens with our modified rules, we plan to evaluate how the addition of these rules affects the intelligibility and naturalness of synthetic fricatives. We already observed that subjects can easily classify the CVs used in the present study as natural or synthetic, even with very short vowels. Finally, we must investigate additional speakers and consider higher frequency cut-offs to determine whether the phenomena we observed are typical.

5. ACKNOWLEDGEMENTS

This work was supported in part by a grant from the National Institute of Neurological and Communicative Disorders and Stroke and the National Science Foundation, no. DC0075.

6. REFERENCES

[1] Badin, P. and Fant, G. (1989), "Fricative production modelling: aerodynamic and acoustic data. *Eurospeech 89 Conference, Paris. Vol. 2.*

[2] Behrens, S. J. & Blumstein, S. E. (1988), "Acoustic characteristics of English voiceless fricatives: A descriptive analysis", *J. Phonetics , 16,* 295-298.

[3] Harris, K. S., (1958), "Cues for the discrimination of American English fricatives in spoken syllables. *Lang. Speech 1, Part 1,* 1-7.

[4] Klatt, D. H. (Unpublished manuscript), "Fricative consonants"

[5] Klatt , D. H. and Klatt, L. C. (1990), "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *J. Acoust. Soc. Am, 87(2,* 820-857.

[6] Stevens, K. N. (1987), "Interaction between acoustic sources and vocal-tract configurations for consonants", *In Proceedings of 11th International Conference of Phonetic Sciences, Estonia, Vol. 3,* 385-389.