The Perception of Fricative Peaks and Noise Bands

Jonathan Goodacre and Yoshitaka Nakajima

Department of Acoustic Design, Kyushu University

Abstract Recent work on the identification and perception of fricatives has focused on the use by listeners of spectral moments derived from the whole spectrum and there appears to be no work in the literature on the use of prominent spectral peaks. In this study, we map the response of a single listener to narrow bands of noise that “mimic” the spectral peaks of English voiceless fricatives. The stimuli are based on the critical-band rate scale (Zwicker and Fastl, 1990) which divides the audible frequency range up to 15500 Hz into 24 abutting critical bands. The results suggest that listeners have knowledge that enables them to connect a narrow-band spectral peak with a particular fricative consonant. We demonstrate that such knowledge, particularly in conjunction with a normalization metric that takes account of an individual speaker’s vocal tract characteristics (F0 of the vowel following the fricative), could be used to good effect, particularly in noisy conditions which impair the use of the whole spectrum.

http://www.jstage.jst.go.jp/browse/jpa

Keywords: speech perception, fricatives, spectral peak, noise band, acoustic design

Introduction

This paper examines the use of spectral peaks in the perception of voiceless fricatives. In recent years, the prevailing interest in the speech perception literature has been in the use of spectral moments; mean, variance, skew, and kurtosis derived from the whole spectrum to classify and discriminate between the different fricative consonants (Forrest et al., 1988; Jongman et al., 2000). Interestingly, there appears to be no work in the literature on the use of single spectral peaks in fricative speech perception. This lack of interest is understandable for a number of reasons. Firstly, the voiceless consonants /f/ and /th/ are spectrally rather flat and often have “no major spectral prominence” (Stevens, 1997), although there is some evidence that the peak of /f/ tends to be at a lower frequency whereas that of /th/ tends to be at a higher frequency. Goodacre and Takagi (2004) found that the peak of /th/ was usually located above 5000 Hz while the peak of /f/ was usually located below 3000 Hz and often below 1500 Hz. This finding is consistent with phonetic/phonological characterizations (Ladefoged, 1997) under which /f/ is classified as “+grave” denoting the existence of more energy in lower frequency bands. In contrast to the relatively flat spectra of /f/ and /th/, the consonants /s/ and /sh/ have prominent spectral peaks. For any individual speaker, the highest-amplitude peak of /s/ is located at a higher frequency than the highest amplitude peak of /sh/. However, problematically, for two speakers with quite different vocal tract lengths (e.g., a large man with a low-pitched voice and a woman with a high-pitched voice) identification of these two fricatives on the basis of spectral peak frequency alone would be difficult. For example, the spectral peak of the woman’s /sh/ might be at a similar or even higher frequency than the peak of the man’s /s/. From the above, we can see that although the spectral peak of a fricative noise in isolation has limited information value, there should be more reliable information when the frequency of the spectral peak is known to come from a particular vocal tract / speaker—for example a man with a low F0 or a child with a high F0. Furthermore, the use of just the spectral peak of fricatives would be particularly advantageous in adverse noise conditions where only prominent spectral peaks are “visible” above the interfering masking noise. Such a strategy would be particularly useful for the contrast /sh/ and /s/ since the two consonants carry a high linguistic load in that they differentiate over a thousand pairs of English words (e.g., sow, show, sheet, sea). Furthermore the fact that the absolute level of the peak of /sh/ and /s/ is greater than the level of /f/, /th/ and /f/ makes these two consonants more robust against the effects of masking noise. From the above, we can see that speech perception mechanisms might be able to make use of the peak part of the spectrum (not only of measurements based on the whole spectrum).

The aim of this paper is twofold. First, to examine the discriminability of the fricatives /sh/ and /s/ based on a speaker-dependent metric M which is computed from the ratio of the frequency of the most prominent spectral peak to the fundamental frequency of the following vowel. The focus will be on the /sh/ versus /s/ contrast since this contrast carries a high linguistic load in English and, as also mentioned above, the peaks of these two fricatives are the most prominent
amongst fricative consonants and hence more resistant to noise interference in adverse listening conditions. This first step will establish whether a robust cue exists in the acoustic signal. The second aim is to map the response of a listener to narrow bands of noise that “mimic” the spectral peaks of fricatives. The stimuli are based on the critical-band rate scale (Zwicker and Fastl, 1990) which divides the audible frequency range up to 15500 Hz into 24 abutting critical bands. This scale takes account of known physiological properties of the auditory system and psychoacoustic findings—such as the fact that auditory filters (in absolute terms) become wider with increasing center frequency. Under this scale, the first band ranges from 0 to 100 Hz, the next from 100 to 200 Hz. The 16th band ranges from 3150 to 3700 Hz and the final band ranges from 12000 to 15500 Hz. Using this type of stimulus enables us to attempt to connect the speech perception mechanisms of listeners to fundamental properties of the auditory system. The purpose here is a “proof of concept” exercise which will assess the viability of using a particular type of stimulus as a research tool.

**Acoustic Measurements**

Tokens came from the Takagi-Goodacre corpus (Goodacre and Takagi, 2004); a corpus of 6 British and American speakers (two males and four females aged from 20 to 22 years old). All speakers had read single-syllable words from a computer screen into an omnidirectional Sony ECM-M957 microphone. The recording device was a Sony DAT with a sampling rate of 44100 Hz. The recordings were transferred from DAT to a PC using SoundForge software. Spectral analyses of individual tokens were made using the MultiSpeech speech Analysis package (Kay Elemetrics, Corp.). Fricative calculations are based on the spectral points below 11000 Hz using a 40 ms Hamming window and generated by means of FFT centered on the point of 2/3 rds duration.

This protocol was chosen so as to be in line with recent approaches to fricative measurements (e.g., Jongman et al., 2000). The point of 2/3 rds duration was chosen because it is known that salient fricative peaks are more likely to emerge in the middle and end of fricative noise (Behrens and Blumstein, 1988). The F0 of vowels was measured by using a zero-crossing method to determine the beginning of the second and fifth periods following vowel onset. The total duration of these three cycles was used to derive the fundamental frequency.

The data are displayed in Table 1. As we can see, for /sh/ M ranges from a minimum value of 24 to a maximum of 32. For /s/, M ranges from a minimum of 36 to a maximum of 48. The data indicate that the metric is completely reliable in discriminating between /sh/ and /s/.

**Interim Summary**

The acoustic measurements reveal that the highest amplitude spectral peak of the fricatives /sh/ and /s/, when used in conjunction with a speaker-dependent normalization (in this case the F0 of the following vowel), provides perfect discrimination between the two consonants. This knowledge is immediately useful in several ways. For example, our metric could be implemented as a strategy in a speech-recognition system. The metric could also be used as an index of the acceptability of syllables produced by synthetic or non-native speech. Of course, the existence of a reliable acoustic cue does not guarantee that human listeners make any use of that cue in their speech perception processes. Furthermore, in order to proceed to productive basic research in this area, it is necessary to develop suitable stimuli that can be used to represent the spectral peak of fricatives. To this end, we now report data from a single listener in which we evaluate the use of narrow bands of noise to mimic fricative peaks.

**Experimental Methods**

A native speaker of English (the first author) participated in the experiment. He reported no medical history of hearing problems. Testing was carried out in a sound-proof room. Stimuli were generated by the computer (44100 Hz sampling rate) using the J programming language and passed through a D/A converter and an anti-aliasing filter with a cut-off frequency of 16000 Hz. The output of the filter was passed through a Stax amplifier and the subject listened monaurally (right ear) through Stax headphones. As mentioned in the Introduction, the stimuli were based on the critical-band rate scale which divides the audible frequency range up to 15500 Hz into 24 abutting critical bands. There were thus 24 stimuli. Each stimulus was composed of 100 components randomly selected to cover the frequency band on a logarithmic scale. The stimuli were presented in blocks of 24 trials with presentation order randomized within a block. After

---

**Table 1** Measurements of highest amplitude spectral point and following vowel F0

<table>
<thead>
<tr>
<th>Word</th>
<th>Peak</th>
<th>F0</th>
<th>M</th>
<th>Word</th>
<th>Peak</th>
<th>F0</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shin</td>
<td>2627</td>
<td>97</td>
<td>27</td>
<td>Sin</td>
<td>3962</td>
<td>99</td>
<td>40</td>
</tr>
<tr>
<td>Ship</td>
<td>2670</td>
<td>103</td>
<td>26</td>
<td>Sip</td>
<td>4335</td>
<td>105</td>
<td>42</td>
</tr>
<tr>
<td>Short</td>
<td>3273</td>
<td>105</td>
<td>31</td>
<td>Sought</td>
<td>3919</td>
<td>103</td>
<td>38</td>
</tr>
<tr>
<td>Shin</td>
<td>2454</td>
<td>102</td>
<td>24</td>
<td>Sin</td>
<td>4220</td>
<td>106</td>
<td>40</td>
</tr>
<tr>
<td>Ship</td>
<td>2497</td>
<td>106</td>
<td>24</td>
<td>Sip</td>
<td>3876</td>
<td>107</td>
<td>36</td>
</tr>
<tr>
<td>Short</td>
<td>2971</td>
<td>103</td>
<td>29</td>
<td>Sought</td>
<td>3919</td>
<td>100</td>
<td>39</td>
</tr>
<tr>
<td>Shin</td>
<td>4651</td>
<td>178</td>
<td>26</td>
<td>Sin</td>
<td>7062</td>
<td>182</td>
<td>39</td>
</tr>
<tr>
<td>Ship</td>
<td>5254</td>
<td>174</td>
<td>30</td>
<td>Sip</td>
<td>7536</td>
<td>180</td>
<td>42</td>
</tr>
<tr>
<td>Short</td>
<td>5297</td>
<td>181</td>
<td>29</td>
<td>Sought</td>
<td>6503</td>
<td>176</td>
<td>37</td>
</tr>
<tr>
<td>Shin</td>
<td>5383</td>
<td>168</td>
<td>32</td>
<td>Sin</td>
<td>8139</td>
<td>170</td>
<td>48</td>
</tr>
<tr>
<td>Ship</td>
<td>4300</td>
<td>172</td>
<td>25</td>
<td>Sip</td>
<td>6976</td>
<td>175</td>
<td>40</td>
</tr>
<tr>
<td>Short</td>
<td>4785</td>
<td>165</td>
<td>29</td>
<td>Sought</td>
<td>7192</td>
<td>171</td>
<td>42</td>
</tr>
</tbody>
</table>

**Average M** 25.7 40
each trial, the subject had the option to replay the stimulus. The subject performed several practice blocks before undertaking four experimental blocks which were to be used for data analysis. On the basis of these practice blocks, the subject adjusted the presentation level (loudness) to a subjectively “comfortable level just below the point which would start to become slightly loud” by adjusting the output of the Stax amplifier. The subject recorded his response on a response sheet. The task, for each trial, was to allocate points out of a total of ten to the perceptual impression evoked by the stimulus. For example, if a stimulus elicited a strong “sh” sensation together with a slight sensation of “s”, then the subject should record a response of 7 and 3 or 8 and 2 for “sh” and “s” respectively. If a stimulus gave equally strong sensations of “f” and “th” then a 5:5 allocation for these two consonants would be appropriate. Overall, there were six response categories; i.e., the five English voiceless fricatives (/h/, /θ/, /ʃ/, /s/, /θ/) and “N” for an indeterminate noise sensation.

Results and Discussion

Table 2 shows the data for the listener. The vertical gives the number of the noise band on the critical-band Rate scale. The response options are written horizontally. The response scores are the average of the 4 trials per stimulus. Where a particular score exceeds 5 then it is highlighted in bold. As we can see, low-frequency noise bands in the range up to the 11th band (center frequency of 1370 Hz) evoke a predominant “th” response with a lesser “f” response. Weak “th” responses are only evoked for stimuli from the high-frequency bands 20 to 23. The respective frequency ranges for “f” and “th” responses seem in line with the acoustic measurements reported in the Introduction (Goodacre and Takagi, 2004). Moderately strong /ʃ/ responses begin with the 12 th band with the transition to /s/ occurring between the 17th and 18th bands (i.e.4000 Hz). The /s/ region extends from the 19th to the 22nd bands. The values for center points of the ranges formed by dominant responses for “sh” and “s” are 2500 Hz and 5300 Hz, respectively. These values are slightly shifted upwards from what would be predicted by a mapping based on the average fundamental frequency of 147 Hz for adult men and women (Hillenbrand et al., 1995) and the acoustic measurements from Section II above. However the transition point between the two fricatives is slightly shifted down from that expected on the basis of the measurements. Thus, broadly speaking the values for “sh” and “s” have some compatibility with the acoustic measurements of section II.

Overall the results indicate that the listener has an internal mapping that enables him to connect a narrow-band noise mimicking a spectral peak with a particular fricative consonant. We can suppose that such knowledge, particularly in conjunction with knowledge of an individual speaker’s vocal tract characteristics (such as vowel F0), could be used to good effect in noisy conditions that impair the use of the whole spectrum. In masking noise conditions that afforded only brief “glimpses” of a fricative’s prominent spectral peak, such information would still be highly informative.

In conclusion, this paper has presented two connected pieces of work that contribute to our ongoing project on the use of fricative peaks in speech perception. We have shown that a metric using the highest amplitude peak of the fricative and the F0 of the following vowel can discriminate perfectly between the consonants /ʃ/ and /s/. Following on from this, we have shown that it is possible to map the responses of a listener to narrow-band noises that “mimic” the spectral peaks of fricatives. It should be noted that the results are obtained from a single experienced and well-trained listener and that the strongest responses tend to score around 6 to 7. It may be that less experienced listeners would give less clear-cut results. Furthermore it remains to be seen whether superior stimuli for “mimicking” fricative peaks, which elicit yet stronger responses, can be developed. For example, stimuli spanning several critical bands and declining by around 10 dB in amplitude from the central peak to the edges of the frequency band would be more similar in their spectral profile to actual fricative productions. That said, the strong point of the current stimulus-design, based on the critical-band rate scale, is that it enables us to relate phoneme perception to a fundamental property of the auditory system.

The above limitations aside, this work has established that, in principle, listeners have prototypical knowledge of fricative peak location that may be available for use by speech
perception processes. Furthermore we have shown that for the /sh/ versus /s/ contrast, a speaker-dependent metric which uses the higher amplitude peak normalized by the F0 of the following vowel provides robust information. Future work should examine how listeners use their knowledge when they are presented with stimuli which contain both fricative like narrow noise-bands and accompanying synthetic vowels. An obvious continuation is to use narrow noise band stimuli spanning the transition region of “sh” and “s” of individual subjects and to evaluate the extent to which variation in vowel fundamental frequency can shift the response function for the two fricatives.

From a future applied perspective, increased knowledge of the use of fricative peaks in human speech perception and of variation in individual “fricative-peak maps” may inform the better design of hearing aids, human telecommunication devices and synthetic speech regarding the “safe” cut-off frequency for high-frequency fricative information. Finally, implementation of the metric $M$ in machine speech recognition might afford extra robustness in adverse listening conditions where only prominent spectral peaks are sometimes available as input above the background noise.

Acknowledgments This work was a part of the 21st Century COE program from the Japanese Ministry of Education, entitled “Design of artificial environments on the basis of human sensibility”.

References


Received: September 6, 2004
Accepted: November 11, 2004
Correspondence to: Jonathan Goodacre, Department of Acoustic Design, Kyushu University, 4–9–1 Shiobaru, Minami-ku, Fukuoka 815–8540, Japan
Phone: +81–92–553–4558
e-mail: kampnaka@yahoo.com