EFFECT OF FRICATION DURATION AND FORMANT TRANSITIONS ON THE PERCEPTION OF FRICATIVES IN VCV UTTERANCES

K. S. Nataraj, Prem C. Pandey, Hirak Dasgupta

Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India
natarajks@ee.iitb.ac.in, pcpandey@ee.iitb.ac.in, hirakdgpt@ee.iitb.ac.in

ABSTRACT

Place perception for fricatives is reported to be determined by the spectral characteristics of the frication segment and also by the formant transitions in case of non-sibilants. A study is conducted to investigate the relative importance of the transitions and the frication duration on the perception of the unvoiced fricatives /f, s, sh/. Listening tests for consonant identification were conducted using vowel-fricative-vowel sequences, generated using the vowel and transition segments extracted from natural utterances and the frication segments synthesized with duration of 30–300 ms. The test results showed that the sequences with frication duration of 50 ms and larger were identified as fricatives, with the place determined by a combination of the transition and frication segments. The place perception for sequences with transition corresponding to /f/ was not affected by the frication segments, while that for sequences with transitions corresponding to /s/ and /sh/ was dominated by the frication segments. The findings indicate that estimation of the place of articulation of fricatives may be improved by combining the information from the transition and frication segments.

Index Terms— fricatives, frication noise, formant transitions, place of articulation

1. INTRODUCTION

Fricatives are produced when air is forced through a narrow constriction of the oral cavity to create turbulence [1], [2]. Based on the place of articulation (place of maximum constriction in the oral cavity), the English fricatives are grouped into four classes: (i) labiodental (/f/ as in "fine", /v/ as in "vine"), (ii) linguodental (/θ/ as in "thing", /ð/ as in "then"), (iii) alveolar (/s/ as in "sue", /z/ as in "zoo"), (iv) palatal (/ʃ/ as in "shoe", /ʒ/ as in "measure").

Earlier studies have investigated the importance of several acoustic characteristics of fricatives including spectral peak location, spectral moments, spectral slope, duration, and energy for characterizing the place of articulation of fricatives [3]–[9]. Li *et al*. [5] used '3-dimensional deep search' method, using speech material from three male and three female speakers, to isolate the spectral cue for the perception of American English fricatives. They reported the spectral cue regions for the labial, alveolar, and palatal fricatives as 0.6–2.2, 3.6–8, and 1.4–4.2 kHz, respectively. Ali *et al*. [6] used a normalized spectral slope parameter to characterize /f, v/. In an acoustic study of fricatives produced by children of approximately 5 years, Nissen and Fox [8] found slope and variance of the spectrum to be important for their identification. Zharkova *et al.* [9] studied the variability in tongue shapes obtained from ultrasound imaging and spectral characteristics of the fricatives /s, ʃ/ produced by children of 7–11 years. The tongue shape was characterized by amount of tongue excursion in relation to the ends of the tongue curve (termed as bunching) and the location of the most bunched part of the tongue measured from the end of tongue curve. They reported that spectral centroid and tongue shape were different for the two fricatives and the discriminability increased with the age. Thus, these studies indicate that the place for fricatives is related to the spectral characteristics of the frication segment. They further indicate that the alveolars /s, z/ have energy concentrated in the high-frequency region, while the labiodentals /f, v/ have it in the low-frequency region. However, the spectral cues for different places have significant overlap and utterance-to-utterance variability. An analysis of several fricatives in the XRMB database [10] showed considerable variations in the relations between spectral characteristics and place. Figure 1(a) shows an example of the alveolar fricative /s/ with the spectral energy distributed uniformly instead of being concentrated in the high-frequency region and Figure 1(b) shows an example of the labiodental fricative /f/ with the spectral energy evenly distributed instead of being concentrated in the low-frequency region. These utterances having similar frication spectra are perceived distinctly and indicate that the characteristics of the frication spectra alone may not convey the place.

It has been established that the formant transitions in the vocalic segments preceding or following the stop closure provide an important cue for perception of the place of stops [11]. Harris [12] investigated importance of vocalic segment on perception of the place of fricatives, using the consonant-vowel (CV) utterances from one male speaker with interchange of the vocalic and frication segment of different CV utterances. The study concluded that the vocalic segment provides an important cue for perception of the non-sibilants /f, θ/, while perception of the sibilants /s, ʃ/ depends primarily on the frication segment. Wagner *et al.* [13] used pseudo words involving misleading formant transitions for /s/ and /f/ preceded and followed by /a, i, u/ spoken by Dutch and Spanish speakers. Listening tests for phoneme identification with the native speakers of Dutch, English, German, Polish, and Spanish as the listeners, showed that the formant transitions provided important cues for English, Polish, and Spanish listeners as these languages include spectrally similar fricatives. To study the importance of the frication duration on the place identification, Jongman [14] conducted listening tests using CV utterances comprising the fricatives /f, v, θ, ð, s, z, ʃ, ʒ/ and the vowels /a, i, u/ spoken by a male speaker, with the utterances edited to include 20–70 ms of the frication and the utterances with the entire frication. The test results indicated that 30–50 ms of frication was

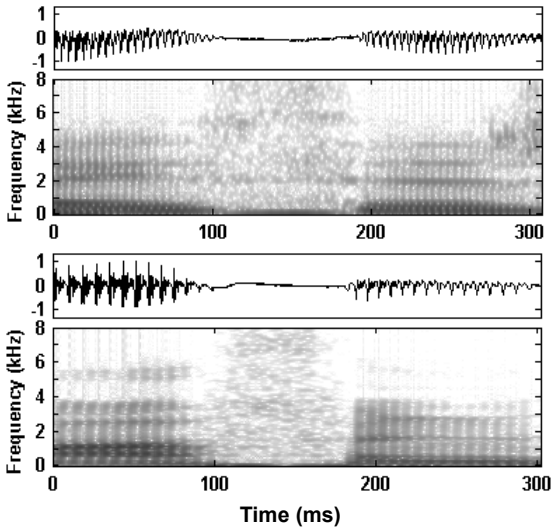

Figure 1: Examples of fricatives with frication spectra deviating from the ones commonly reported in the literature: Waveform and spectrogram of (a) /ɛsʌ/ (speaker JW62, Task 78), (b) /afʌ/ (speaker JW44, Task 24), from the XRMB database.

required for identification of /f, v, s, z, ʃ, ʒ/, while the entire frication was required for identification of /θ, ð/. A decrease in the frication duration affected the place identification more than the manner and voicing identifications.

Earlier studies have examined the effect of the vocalic segment and the frication duration on the perception of the place of articulation of fricatives separately. However, the effect of the vocalic segment may be less on the utterances with frication of long duration. Thus it may be useful to investigate the effect of the vocalic segment on the perception of the place of articulation of fricatives with different frication durations. We present a study to investigate the relative importance of the transitions and the frication duration on the perception of the unvoiced fricatives /f, s, ʃ/. The material and method are described in the second section, the results of the experiments are presented in the third section, followed by the conclusion in the last section.

2. MATERIAL AND METHOD

2.1. Speech material

A set of VCV utterances involving unvoiced fricatives /f, s, ʃ/ with vowels /a, i, u/ were recorded at 16 kHz from two male and two female speakers. All the speakers had studied English as their first or second language and were able to pronounce all the fricatives with acceptable quality and intelligibility. The fricative /θ/ was not included as many in the group of speakers and subjects had difficulty in discriminating it from a stop with nearly the same place in their first language. Based on informal listening, VCV utterances with vowel /a/ from one male speaker were selected as the speech material for the investigation. To study the effect of formant transitions and frication duration on the perception of the place of fricatives, the sequences were generated with different frication durations. Multiple utterances with the same vowel-fricative-vowel have significant utterance-to-utterance and duration-related variations in the frication spectrum. To avoid the effects of these variations on the results of the listening test, synthesized frication segments were used in the test sequences.

The VCV utterances were edited to extract the vowel-fricative and fricative-vowel segments. The frication segment for each fricative was synthesized by filtering white noise using an FIR filter with its frequency response approximating the averaged spectrum of the frication segment of the natural utterance. The RMS value of the synthesized fricative was made equal to that of the natural fricative. The mel-cepstral distances [15] between the synthetic and naturally produced frications for the fricatives /f/, /s/, and /ʃ/ were 3.8, 5.6, and 5.2 dB, respectively. The corresponding mean values of the distances between multiple natural utterances of these fricatives by the same speaker were 10.2, 18.3, and 10.2 dB. Thus the synthetic frications may be considered as similar to the naturally produced frications. Each VCV sequence for the test material was generated by concatenating the extracted VC segment, the synthesized frication segment, and the extracted CV segment. To avoid a perceptible discontinuity in the concatenation, the amplitude envelope of the frication was multiplied with a trapezoidal window with 20-ms rising and falling sub-segments, as shorter rise and fall resulted in perceptible distortion and longer rise and fall decreased the steady part of the frication. No perceptible distortion was noticeable in the resulting VCV sequences.

To study the effect of frication duration, the VCV sequences were generated using frication duration varying from 30 ms to 300 ms. It has been reported that a frication duration of 70 ms is sufficient for correct identification of the fricative [14]. The sequences with longer frication duration were also included to study the effect of the vocalic segment on the sequences with long frication duration. Sequences with a total of eight durations (30, 50, 70, 90, 120, 150, 200, 300 ms) along with three transitions, and three fricatives, resulting in 72 VCV sequences, were generated for the listening test.

2.2. Experimental method

Ten normal-hearing students (eight males and two females, 22–35 years) served as the subjects for the listening test. All the subjects had studied English as their first or second language and were able to perceive the differences between all the sounds used in the study. The test was administered using a graphical user interface (GUI) on a PC-based automated experimental setup for consonant identification. The stimuli comprising the 72 VCV sequences, as described earlier, were presented in a randomized order using a Sennheiser HD 202 headphone at the most comfortable level as selected by the subject. The task involved listening to the stimulus and responding by clicking one of the seven response buttons, including three buttons for the fricatives /f, ʃ, s/, two buttons for the stops /p, t/, one button for the affricate /tʃ/, and one button for 'none of above'. The subject was permitted to play the sound using the 'play' button on the GUI before responding. A subject was allowed to pause the listening test and resume the test later.

3. RESULTS

3.1. Results from listening test

There were 720 responses (10 subjects, 72 stimuli). The results of the listening tests are given in Figure 2, with each bar representing the responses to one of the 72 stimuli.

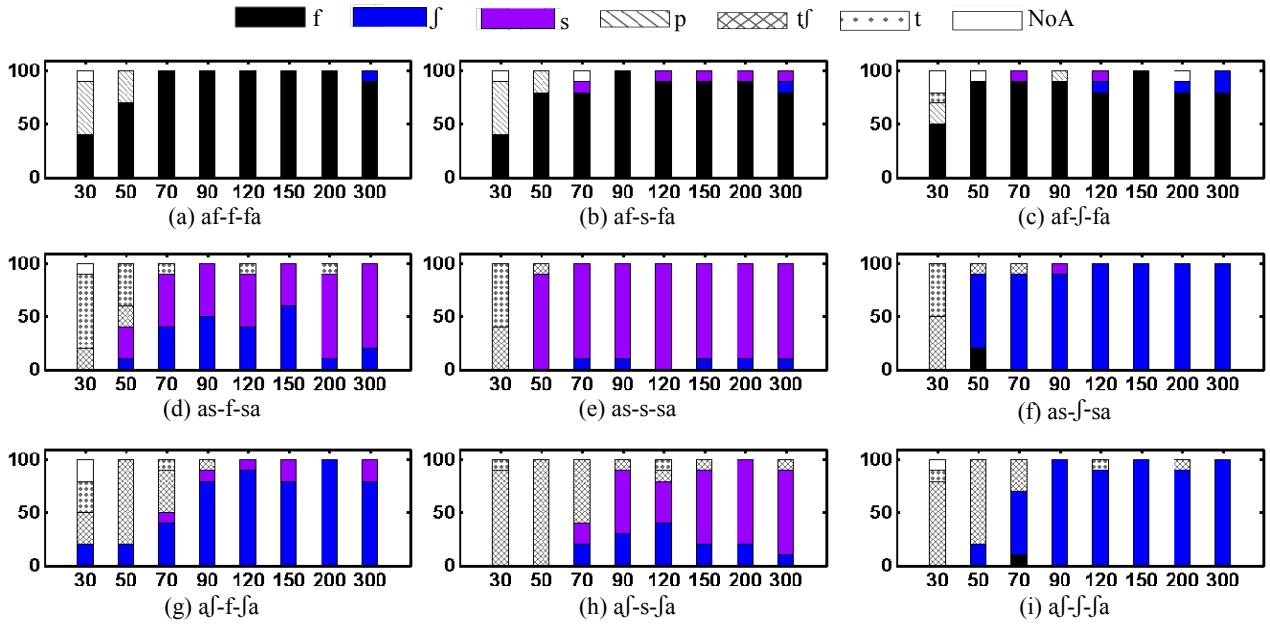

Figure 2: Consonant identification scores for the test sequences with different transition and frication combinations (e.g. 'af-s-fa' being the test sequence with the VC and CV transition segments from the recorded utterance /afa/ and the synthesized frication corresponding to /s/) and different frication durations: seven response scores (%) as shaded bars along the y axis and duration in ms along the x axis.

The first row (a, b, c) of Figure 2 shows the results for test sequences generated using the VC and CV segments corresponding to the utterance /afa/. For frication duration of 30 ms, the sequences were perceived either as stop /p/ (having the place close to that of /f/) or as fricative /f/, indicating that the transition segments provided the place cue even though the frication was short. However, the responses for these test sequences were 50% or less, indicating the difficulty in consonant identification for short-duration frication. For frication durations of 50 ms and longer, most of the test sequences were perceived as /f/ and the responses were 70% or higher irrespective of the spectral characteristics of the frication. These results indicate that the perception of the test sequences with the transition segments corresponding to /f/ was not affected by the spectral characteristics of the frication. It is interesting to note that the effect of transition in this case is not reduced by an increase in the frication duration.

The second row (d, e, f) of Figure 2 shows the results for test sequences generated using the VC and CV segments corresponding to sequence /asa/. For frication duration of 30 ms, the test sequences were perceived as stop /t/ or affricate/t͡ʃ/. For larger frication durations, the test sequences with the frication segment corresponding to /s/ were identified as /s/ with responses of 70% or higher. For frication duration of 70–300 ms, the test sequences with frication segment corresponding to /f/ were identified as /s/ or /ʃ/ indicating that the transition segments dominated the perception. The test sequences with the frication segments corresponding to /ʃ/ were identified as /ʃ/, indicating that frication corresponding to /ʃ/ dominated the place perception.

The third row (g, h, i) of Figure 2 shows the results for test sequences generated using the VC and CV segments corresponding to sequence /aʃa/. For frication durations of 30 and 50 ms, the test sequences were perceived as affricate /t͡ʃ/ (the place corresponding to the transition segments). The responses for many of these sequences were 70% or higher, indicating that the transition segments dominated the place perception. For frication durations of 70–300 ms, the sequences were perceived as /ʃ/ if the frication segments corresponded to /ʃ/ with responses of 70% or higher. For durations of 90 and 120 ms and frication segments corresponding to /s/, the test sequences were identified as /s/ or /ʃ/ with responses of 70% or less, indicating similarity of the transition segments for /s/ and /ʃ/. For duration larger than 120 ms, the sequences with frication corresponding to /s/ were perceived as /s/ with responses of 70% or higher, indicating that the larger frication duration helped in discrimination between /s/ and /ʃ/. These results show that the frication segment dominated the place perception when the transition segments corresponded to /ʃ/ and frication segment corresponded to /s/. However, when the frication corresponding to /f/ was combined with the transitions corresponding to /ʃ/, the test sequences were identified as /ʃ/ with responses of 70% or higher, indicating that the transition segments dominated the place perception.

3.2. Comparison with earlier reported investigations

Investigations similar to the current one have been reported by Harris [12], Wagner *et al.* [13], and Jongman [14] as described earlier. Harris [12] concluded that perception of the fricatives /s, ʃ/ depends on the frication part alone, irrespective of the transition segment. We observed that the perception of frication corresponding to /s, ʃ/ paired with the transitions corresponding to /f/ is dominated by the transition segment. It may be noted that the current study uses vowel /a/, while the results for this vowel are not reported in [12]. Results of study by Wagner *et al.* [13] are similar to our study for /f, s/, but our study also has results for /ʃ/ and the effect of change in frication duration and the change in transition segment together was not investigated in [13]. In the study by Jongman [14], the effect of the frication duration was studied without interchanging the transition segments and thus it did not investigate the importance of transition segments on the fricative perception as in the current study.

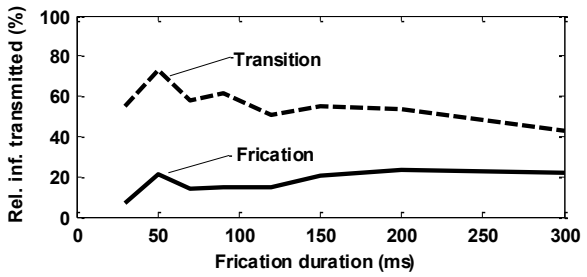

Figure 3: Relative information transmitted (%) for transition and frication features as a function of frication duration.

3.3. Information transmission analysis

Relative information transmission analysis has been used in many studies to analyze the stimulus-response confusion matrices of the listening tests for quantifying the contributions of different input features for consonant identification [14], [16]–[19]. As described by Miller and Nicely [16], the relative information transmitted between the stimulus set **x** and response set **y** is given as

$$I_{rel}(\mathbf{x},\mathbf{y}) = \frac{\sum_{i,j} p(x_i,y_j)\,\log\left[p(x_i)p(y_j)/\,p(x_i,y_j)\right]}{\sum_{i} p(x_i)\,\log\left[p(x_i)\right]} \quad (1)$$

where $p(x_i)$ is the probability of the stimulus x_i, $p(y_j)$ is the probability of the response y_j, and $p(x_i, y_j)$ is the joint probability of the stimulus x_i and the response y_j. For the confusion matrices of our listening test, the relative information transmitted was calculated for each frication duration, with the response set comprising the seven responses (/f, s, ʃ, p, t, ʧ, none of above/) and two features: (i) frication with the stimulus set comprising the three frications (/f, s, ʃ/) and (ii) transition with the stimulus set comprising the three transitions (/af, as, aʃ/).

Figure 3 shows the relative information transmitted as a function of duration for the transition and frication features. It is observed that the transition feature provides more information compared to the frication feature for all durations as transition dominates the perception in most of the cases. This may be due to the experimental conditions, as there were more conditions when the center frication and formant transition did not match and the formant transitions dominated the perception during these mis-matches. The sequences with 30 ms were difficult for consonant identification. As the frication duration increases above 50 ms, the information transmitted by the frication feature increases and that by the transition feature decreases. This variation is due to the cases for which frication part dominates the perception with transition corresponding to the /s, ʃ/ cases and the frication corresponding to the /s, ʃ/ cases.

3.2. Spectral analysis of transition segments

Spectrograms during the vowel-consonant and consonant-vowel transitions of multiple VCV utterances with vowel /a/ and fricatives /f, s, ʃ/ from one speaker were examined for spectral cues for place during misleading transitions. The variation of the second formant frequency during transitions was observed to be consistently related to the fricatives. During the VC transition for /f/, the second formant decreased by about 200 Hz. During the VC transition for /s/ and /ʃ/, the second formant increased by about 100 Hz and 200 Hz, respectively. Thus, the second formant transitions for the labiodental fricative differed from that for the alveolar and palatal fricatives. The similarity in the CV transitions for the alveolar and palatal fricatives may be the reason that perception of test sequences with these fricatives was dominated by the frication segment. Formant transitions during the CV segments were complementary to those during the VC segments.

4. CONCLUSIONS

Investigations were carried out on the perception of the unvoiced fricatives /f, s, ʃ/ in vowel-consonant-vowel sequences with natural transition segments and synthesized frication segments with durations of 30–300 ms. Results of the listening test showed that the test sequences having frication duration less than 50 ms were difficult for identification and were identified as stops with the place determined by the transition for most of the sequences. The sequences with larger frication durations were identified as fricatives, with the place determined by a combination of the transition and frication segments. The place perception for sequences with transition corresponding to /f/ was not affected by the frication segments. For sequences with transitions corresponding to /s/ and /ʃ/, the place perception was dominated by the frication segments if the frication segment corresponded to /s/ or /ʃ/. However, the transition segments dominated the place perception for the test sequences with the frication segments corresponding to /f/ and transition segments corresponding to /s/ or /ʃ/. Thus it may be concluded that the transition segments dominate the place perception in case of a mismatch between the transition and frication segments. The results further indicate that estimation of the place of articulation of fricatives may be improved by combining the place information obtained from the transition and frication segments. The investigation needs to be extended for test sequences with other fricatives, different vowel contexts, and using speech material from male and female speakers.

5. ACKNOWLEDGEMENTS

The research is supported by the project "Visual Speech Training System for the Hearing Impaired" sponsored by the Ministry of Electronics & Information Technology, Government of India. Work of K. S. Nataraj was supported in part by Bharti Centre for Communication in IIT Bombay.

6. REFERENCES

[1] P. Ladefoged, *A Course in Phonetics,* 2nd ed. New York, USA: Harcourt Brace Jovanovich, 1982.

[2] D. O'Shaughnessy, *Speech Communication: Human and Machine*, 2nd ed. Piscataway, NJ, USA: IEEE Press, 2000.

[3] G. W. Hughes and M. Halle, "Spectral properties of fricative consonants," *J. Acoust. Soc. Am.*, vol. 28, pp. 303–310, 1956.

[4] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives," *J. Acoust. Soc. Am.*, vol. 108, pp. 1252–1263, 2000.

[5] F. Li, A. Menon, and J. B. Allen, "A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise," *J. Acoust. Soc. Am.*, vol. 132, pp. 2663–2675, 2012.

[6] A. M. A. Ali, J. V. Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic classification of fricatives," *J. Acoust. Soc. Am.*, vol. 109, pp. 2217–2235, 2001.

[7] K. S. Nataraj, P. C. Pandey, and H. Dasgupta, "Estimation of place of articulation of fricatives from spectral characteristics for speech training," in *Proc. Interspeech 2017,* Stockholm, Sweden*,* 2017, pp. 339–343.

[8] S. L. Nissen and R. A. Fox, "Acoustic and spectral characteristics of young children's fricative productions: A developmental perspective," *J. Acoust. Soc. Am.*, vol. 118, pp. 2570–2578, 2005.

[9] N. Zharkova, W. J. Hardcastle, and F. E. Gibbon, "The dynamics of voiceless sibilant fricative production in children between 7 and 13 years old: An ultrasound and acoustic study," *J. Acoust. Soc. Am.*, vol. 144, pp. 1454–1466, 2018.

[10] J. R. Westbury, "X-ray microbeam speech production database user's handbook (version 1.0)," 1994. [Online]. Available: www.haskins.yale.edu/staff/gafos_downloads /ubdbman.pdf

[11] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *J. Acoust. Soc. Am.*, vol. 64, pp. 1358–1368, 1978.

[12] K. S. Harris, "Cues for the discrimination of American English fricatives in spoken syllables," *Lang. and Speech*, vol. 1, pp. 1–7, 1958.

[13] A. Wagner, M. Ernestus, and A. Cutler, "Formant transitions in fricative identification: The role of native fricative inventory," *J. Acoust. Soc. Am.*, vol. 120, pp. 2267–2277, 2006.

[14] A. Jongman, "Duration of fricative noise required for identification of English fricatives," *J. Acoust. Soc. Am.*, vol. 85, pp. 1718–1725, 1989.

[15] T. Toda, A. W. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis," in *Proc. 5th ISCA Speech Synth. Workshop (SSW5)*, Pittsburgh, PA, USA, 2004, pp. 31–36.

[16] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.*, vol. 27, pp. 338–352, 1955.

[17] B. W. Y. Hornsby and T. A. Ricketts ''The effects of compression ratio, signal-to-noise ratio, and level on speech recognition in normal-hearing listeners,'' *J. Acoust. Soc. Am.* vol. 109, pp. 2964–2973, 2001.

[18] L. Xu, C. S. Thompson, and B. E. Pfingst, "Relative contributions of spectral and temporal cues for phoneme recognition," *J. Acoust. Soc. Am*, vol. 117, pp. 3255–3267, 2005.

[19] N. Zhou, L. Xu, and C. Lee, "The effects of frequency-place shift on consonant confusion in cochlear implant simulations," *J. Acoust. Soc. Am.*, vol. 128, pp. 401–409, 2010.