A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise

Feipeng Li[a)]
Center for Language and Speech Processing, Johns Hopkins University, Baltimore, Maryland 21218

Andrea Trevino, Anjali Menon, and Jont B. Allen
Department of Electrical and Computer Engineering, University of Illinois at Urbana–Champaign, Urbana, Illinois 61801

(Received 1 June 2011; revised 23 July 2012; accepted 30 July 2012)

In a previous study on plosives, the 3-Dimensional Deep Search (3DDS) method for the exploration of the necessary and sufficient cues for speech perception was introduced (Li *et al.*, (**2010**). J. Acoust. Soc. Am. **127**(4), 2599–2610). Here, this method is used to isolate the spectral cue regions for perception of the American English fricatives /ʃ, ʒ, s, z, f, v, θ, ð/ in time, frequency, and intensity. The fricatives are analyzed in the context of consonant-vowel utterances, using the vowel /ɑ/. The necessary cues were found to be contained in the frication noise for /ʃ, ʒ, s, z, f, v/. 3DDS analysis isolated the cue regions of /s, z/ between 3.6 and 8 [kHz] and /ʃ, ʒ/ between 1.4 and 4.2 [kHz]. Some utterances were found to contain acoustic components that were unnecessary for correct perception, but caused listeners to hear non-target consonants when the primary cue region was removed; such acoustic components are labeled "conflicting cue regions." The amplitude modulation of the high-frequency frication region by the fundamental *F*0 was found to be a sufficient cue for voicing. Overall, the 3DDS method allows one to analyze the effects of natural speech components without initial assumptions about where perceptual cues lie in time-frequency space or which elements of production they correspond to.
[http://dx.doi.org/10.1121/1.4747008]

PACS number(s): 43.71.Es [TD] Pages: 2663–2675

I. INTRODUCTION

When isolating the cues used for speech perception, a widely recognized problem for analyzing natural speech is the large variability introduced by the speaker (e.g., pitch and rate). Following the 1930–1940 development of the speech "vocoder" at Bell Labs, speech synthesis has been a hallmark of speech perception research. Beginning at Haskins Laboratories in the 1950s (Cooper *et al.*, 1952; Delattre *et al.*, 1955; Bell *et al.*, 1961), almost all of the classical studies have used vocoder speech (Shannon *et al.*, 1995) or speech synthesis methods (Hughes and Halle, 1956; Heinz and Stevens, 1961; Blumstein *et al.*, 1977; Stevens and Blumstein, 1978) as the main way of controlling the variability of speech cues when performing perceptual experiments. A major disadvantage of this method is that one must first make assumptions about the nature of perceptual cues in order to synthesize the target speech stimuli; depending on the accuracy of these assumptions, this can lead to listeners using different cues for recognition than they would for natural speech. Synthesized speech can sound unnatural or have low baseline intelligibility (Delattre *et al.*, 1955; Remez *et al.*, 1981). Many studies analyzed the spectrum of natural speech (Soli, 1981; Baum and Blumstein, 1987; Behrens and Blumstein, 1988; Shadle and Mair, 1996; Jongman *et al.*, 2000) and identified the acoustic cues sufficient for sound identification/discrimination but without verifying them against human psychoacoustic data. While the results characterize the variability of natural speech, it remains uncertain whether those cues are indeed necessary and sufficient for speech perception.

To determine the cues for speech perception, we have developed a new methodology named the *3-Dimensional Deep Search* (3DDS) (Li *et al.*, 2010) that analyzes the perceptual contributions of naturally-produced speech components based on the results of three psychoacoustic experiments. This is paired with a time-frequency model representation, the AI-gram (Régnier and Allen, 2008; Lobdell, 2009; Lobdell *et al.*, 2011), to predict the audibility of acoustic cues as masking noise is introduced.

A. 3DDS and its application to stop consonants

The objective of 3DDS is to measure the significance of speech subcomponents on perception in three dimensions: time, frequency, and signal-to-noise ratio (SNR). Starting in the 1920s, Fletcher and his colleagues used masking noise along with high- and low-pass filtered high-entropy "nonsense" syllables to study the contribution of different frequency bands to speech perception, as a function of SNR (Fletcher and Galt, 1950; French and Steinberg, 1947; Allen, 1994). These classic studies led to the *articulation*

a)Author to whom correspondence should be addressed. Electronic mail: fli12@jhmi.edu

index (AI) model of speech intelligibility (ANSI S3.5, 1969). Based on the AI, Lobdell and Allen developed a computational model denoted the *AI-gram* that simulates the effect of noise masking on audibility (Lobdell, 2009; Lobdell *et al.*, 2011). The AI-gram is an integration of the AI model of speech intelligibility and Fletcher and Galt's critical-band auditory model (i.e., Fletcher and Galt's SNR model of signal detection). Given a speech sound and masking noise, the AI-gram simulates the effect of noise masking and produces an image that predicts the audible speech components along the time and frequency axes. In Miller and Nicely (1955), Wang and Bilger (1973), and Allen (2005), noise masking was used to study consonant confusions. In 1986, Furui used time-truncation experiments to analyze the essential time-waveform components for speech perception. All of these techniques are merged for the 3DDS methodology, which uses three independent psychoacoustic experiments and the AI-gram to evaluate the contribution of speech components to consonant perception.

To isolate the perceptual cue of a consonant-vowel (CV) utterance, the 3DDS method is composed of three independent psychoacoustic experiments that modify the speech along time, frequency, and SNR (see Fig. 1). The naming paradigm for each experiment (TR07, HL07, and MN05) is set up such that the two-digit suffix indicates the year when the experiment was performed. The first experiment (TR07) uses truncation in order to find the location in time or minimum possible duration of the cue region (Li *et al.*, 2010). The second experiment (HL07) is designed to isolate the perceptual cue region in frequency by high- or low-pass filtering the speech at 10 cutoff frequencies that span from 0.25 to 8 [kHz] (Li and Allen, 2009). A third experiment (MN05) assesses the masked threshold (i.e., perceptual robustness to noise) of the speech cue region by masking the speech with white noise at various SNRs (Phatak *et al.*, 2008).

In a previous study, the 3DDS method was used to explore the perceptual cues of stop consonants (Li *et al.*, 2010). It was discovered that natural speech sounds often contain *conflicting cue regions* that lead to confusion when the target-consonant cue region is removed by filtering or masking noise. Through the manipulation of these spectral conflicting cue regions, one consonant can be *morphed* into another or a perceptually weak consonant can be converted into a strong one (Li and Allen, 2011; Kapoor and Allen, 2012). Natural fluctuations in the intensity of the plosive consonant cue regions were found to account for the large variations in the AI (Singh and Allen, 2012). In the present study, the 3DDS method is generalized to fricative consonants.

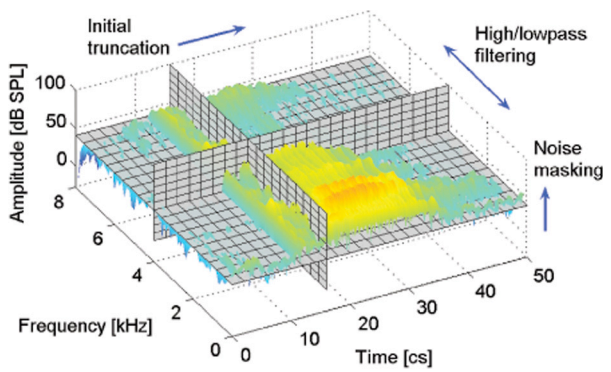

FIG. 1. (Color online) Schematic diagram of 3DDS to characterize the contribution of speech subcomponents to perception as a function of time, frequency, and intensity.

B. Past studies on fricative consonants

Fricative consonants are a major source of perceptual error under noisy conditions, thus they are of special interest. This is true for clearly articulated speech (Miller and Nicely, 1955) as well as maximum entropy CV utterances (Phatak *et al.*, 2008). These studies showed that the non-sibilant fricatives /f, v, θ, ð/ are involved in more than half of the confusions at 12 [dB] SNR in white noise. In contrast, the sibilant alveolars /s, z/, and postalveolars /ʃ, ʒ/ are seldom confused with any other consonants at the same noise level.

Fricative consonants are produced by forcing air through a narrow constriction of the vocal tract above the glottis (Stevens *et al.*, 1992). An early study by Miller and Nicely (1955) observed that the frication noise of the voiced consonants is modulated by *F*0. Miller and Nicely (1955) also noted that the frication regions of /s, ʃ, z, ʒ/ are of longer duration than /f, v, θ, ð/. A consistent difference between "voiced" and "unvoiced" fricatives is the presence of energy below 0.7 [kHz] (Hughes and Halle, 1956) as well as the duration of the frication (Baum and Blumstein, 1987; Stevens *et al.*, 1992). Stevens *et al.* (1992) found that listeners based their voicing judgments of intervocalic fricatives on the time interval duration for which there was no glottal vibration. If this time interval was greater than 6 [cs], the fricative was typically judged as unvoiced (Stevens *et al.*, 1992). When reporting time in our study, the unit centiseconds [cs] is used, as 1 [cs] is a natural time interval in speech perception. For example, an *F*0 of 100 [Hz] has a period of 1 [cs], while relevant perceptual times are always $\geq$1 [cs]. The minimal duration of the frication noise is approximately 3 [cs] for /z/ and 5 [cs] for /f, s, v/ (Jongman, 1988). /θ, ð/ are identified with reasonable accuracy only when at full duration (i.e., no time-truncation) (Jongman, 1988). Although the mean duration of unvoiced fricatives is generally longer than that of the voiced fricatives, the distribution of the two overlap considerably (Baum and Blumstein, 1987).

A number of studies have concluded that /s, z/ are characterized by a strong concentration of frication energy around 4 to 5 [kHz], while /ʃ, ʒ/, pronounced with a longer anterior cavity, have a spectral peak around 2 to 3 [kHz] (Miller and Nicely, 1955; Hughes and Halle, 1956; Heinz and Stevens, 1961; Jongman *et al.*, 2000). Harris (1958) used CV utterances to investigate the relative importance of cues in the frication noise vs the following vocalic portion; utterances were modified by swapping the vocalic portions of different fricatives. Harris found that the cues that discriminated the place of articulation for /s, ʃ, z, ʒ/ were in the frication noise portion, while the place of articulation of /f, θ/ was perceived based on the vocalic portion, although both were perceived as /θ/ when the frication noise was paired with the vocalic portion of /s, ʃ/. Similarly, /ð/ and /v/

were confused with each other when their vocalic portions were swapped. For the voiced fricatives, Harris (1958) noted that the segmentation of frication and vocalic portions may be imprecise, leading to variable results. A later study used cross-spliced "hybrid" speech to find that both frication noise and formant transitions can be used for the distinction of /s/ and /ʃ/ (Whalen, 1981, 1991). Synthesized stimuli with resonant frequencies around 6.5 to 8 [kHz] usually yielded /f/ and /θ/ responses, with /f/ being distinguished from /θ/ on the basis of the second formant transition in the following vowel (Heinz and Stevens, 1961). In contrast, analysis of natural speech from 20 talkers indicates that both /f, v/ and /θ, ð/ display a relatively flat spectrum without any dominating spectral peak (Jongman *et al.*, 2000). In a CV context, the effects of anticipatory coarticulation on the spectral characteristics of /ʃ, ʒ, s, z/ were smallest when the fricatives were followed by the vowel /ɑ/ (Soli, 1981).

Other acoustic cues such as the amplitude of fricative noise (Behrens and Blumstein, 1988; Hedrick and Ohde, 1993) and spectral moments, including skewness and kurtosis (Shadle and Mair, 1996) have been shown to have minimal perceptual significance. Clearly articulated fricatives have perceptual cues shifted toward the high-frequency region (Maniwa and Jongman, 2008).

To summarize the findings of these many past studies, for the sibilant fricatives /ʃ, ʒ, s, z/, the place of articulation is encoded by the spectral distribution of the frication noise. Naturally-produced voicing can be identified by the presence of a low-frequency (<0.7 [kHz]) component, *F*0 modulations of the frication noise, as well as a longer original duration of frication. The relative perceptual roles of these three characteristics of voiced fricatives remain unclear. No conclusive picture is available for the non-sibilant fricatives /f, v, θ, ð/. These many findings result from studies conducted without noise; the addition of noise allows for an information-theoretic analysis of the errors (Miller and Nicely, 1955).

What are the necessary and sufficient perceptual cues of fricative consonants? In this study, we explore this question by using the 3DDS method to analyze perceptual data from a larger past study which gathered data for 16 consonants, including 6 plosives, 8 fricatives, and 2 nasals, with 6 talkers per consonant (Phatak *et al.*, 2008; Li *et al.*, 2010). Isolating the spectral region that contains the necessary and sufficient perceptual cues is equivalent to stating that normal-hearing listeners can correctly perceive the target consonant if (sufficient) and only if (necessary) the cues contained in that region are present. The results for a related analysis on stop consonants are presented in Li *et al.* (2010). For example, in Li *et al.* (2010), the isolated high-frequency burst for /dɑ/ contains all of the timing, frequency, and voicing cues necessary for correct perception; the consonant cannot be perceived correctly without the isolated spectral region. This second component generalizes the 3DDS analysis to the American English fricatives /ʃ, ʒ, s, z, f, v, θ, ð/.

II. METHODS

The basic methodologies of the three perceptual experiments are given next. For additional details about the experimental methods, refer to Li *et al.* (2010), Li and Allen (2009), and Phatak *et al.* (2008).

A. Subjects

In total, 52 listeners were enrolled over 3 studies, of which 17 participated in experiment HL07, 12 participated in experiment TR07 (one participated in both), and 24 participated in experiment MN05. All listeners self-reported no history of speech or hearing disorder. To guarantee that no listeners with hearing loss or other problems were included in this study, any listener with significantly lower performance was excluded from further testing [see Phatak *et al.* (2008) for details]. The first or primary language of all of the listeners is American English, with all but two having been born in the U.S. No significant differences were observed in the consonant scores or confusion of these two listeners, and hence their responses were included. Listeners were paid for their participation. IRB approval was obtained prior to the experiment.

B. Speech stimuli

Sixteen isolated CVs: /p, t, k, b, d, g, s, ʃ, f, v, θ, ð, z, ʒ, m, n /+/ ɑ/ (no carrier phrase) were chosen from the University of Pennsylvania's Linguistic Data Consortium database (LDC-2005S22, aka *the Fletcher AI corpus*) as the common test material for the three experiments. The speech sounds were sampled at 16 [kHz]. Experiment MN05 used 18 talkers for each CV. When extending the psychoacoustic database with HL07 and TR07, six utterances per CV (half male and half female) were selected. In order to explore how cue characteristics contribute to noise robustness, the six utterances were selected such that they were representative of the CVs in terms of confusion patterns (CPs) and score, based on the results of MN05. Specifically, 1/3 of the utterances were selected as high-scoring sounds, and 1/3 of the utterances were low-scoring sounds. Thus, a total of 96 utterances were used (16 CVs × 6 utterances per CV), 48 of which were fricatives and are reported on here. Sounds were presented diotically (both ears) through Sennheiser HD-280 PRO (Wedemark, Germany) circumaural headphones, adjusted in level at the listener's *Most Comfortable Level* for CV utterances in 12 [dB] of white noise, i.e., ≈70 to 75 [dB] sound pressure level. Subjects were allowed to change the sound intensity during the experiment, which was noted in the log files. All experiments were conducted in a single-walled IAC sound-proof booth, situated in a lab with no windows, with the lab outer door shut.

C. Conditions

Experiment TR07 assesses the temporal distribution of speech information (Li *et al.*, 2010). For each utterance, the initial truncation time is set before the beginning of the consonant and the final truncation time is set after the end of the CV transition. The truncation times were chosen such that the duration of the consonant was divided into frames of 0.5, 1, or 2 [cs]. An adaptive strategy was adopted for the calculation of the sample points. The basic idea is to assign more points where the speech perception scores change rapidly (Furui, 1986). Starting from the end of the CV transition and

moving backwards in time, the scheme allocates 8 truncation times (frames) of 0.5 [cs], then 12 frames of 1 [cs], and finally as many 2 [cs] frames as needed until the onset of the consonant is reached. White noise was added following truncation at a SNR of 12 [dB] (based on the unmodified speech sound), matching the control condition of the filtering experiment (HL07).

Experiment HL07 investigates the distribution of speech information in frequency (Li and Allen, 2009). It is composed of 19 filtering conditions, namely one full-band (FB) condition (0.25 to 8 [kHz]), 9 high-pass, and 9 low-pass conditions. The cut-off frequencies were calculated using Greenwood's inverse cochlear map function; the FB frequency range was divided into 12 bands, each having an equal distance along the human basilar membrane. The common high- and low-pass cutoff frequencies were 3678, 2826, 2164, 1649, 1250, 939, and 697 [Hz]. To this we added the cutoff frequencies 6185, 4775 (high-pass) and 509, 363 [Hz] (low-pass). All speech samples were high-pass filtered above 250 [Hz] based on estimates of the frequency importance region observed by Fletcher (Allen, 1994). The filters were implemented as sixth-order elliptic filters having a stop-band attenuation of 60 [dB]. White noise (12 [dB] SNR) was added to the modified speech in order to mask out any residual cues that might still be audible. Note that for most CVs, 12 [dB] SNR does not affect the probability of correct perception (Phatak *et al.*, 2008).

Experiment MN05 (aka MN16R) measures the strength of the perceptual cue region in terms of robustness to white masking noise. Besides a quiet condition, speech sounds were masked at eight different SNRs: −21, −18, −15, −12, −6, 0, 6, and 12 [dB] (Phatak *et al.*, 2008).

We define the probability of correct detection of the target consonant as P_c. The cutoff frequencies of experiment HL07 are denoted f^H (high-pass) and f^L (low-pass). The SNR_{90} is defined as the SNR at which the target consonant has a probability of correct detection of 90% [$P_c(\mathrm{SNR}) = 0.9$].

All three experiments include a common control condition, i.e., full-bandwidth, full-duration speech at 12 [dB] SNR. The recognition scores for this control condition were verified to be consistent across the three experiments.

D. Experimental procedure

The three experiments (TR07, HL07, and MN05) used nearly identical experimental procedures. A MATLAB® program was written for the stimulus presentation and data collection. A mandatory practice session, with feedback, was given at the beginning of each experiment. Speech tokens were randomized across talkers, conditions, and utterances. Following each stimulus presentation, listeners responded by clicking on the button that was labeled with the CV that they perceived. In case the CV was completely masked by the noise, the listener was instructed to click a "Noise Only" button. Frequent (e.g., 20 min) breaks were encouraged to prevent test fatigue. Subjects were allowed to repeat each token up to 3 times, after which the token was pushed to the end of the list. The waveform was played via a SoundBlaster 24 bit sound card in a PC Intel computer, running MATLAB® via Ubuntu Linux.

E. 3DDS procedure

Each of the three experiments provides estimates of different aspects of the necessary perceptual cue region: The critical temporal information, the frequency range, and the intensity. As the listener P_c curves are roughly monotonic (with small amounts of jitter due to random listener error), linear interpolation was used in the analysis of the results. The minimum duration of frication or location in time of the perceptual cue is determined by the truncation time at which the P_c drops below a threshold of 90%. The perceptual cue robustness to noise is defined by the SNR at which the P_c falls below the 90% threshold (SNR_{90}).

The upper and lower frequency boundaries of the perceptual cue region are determined from the high- and low-pass P_c results, respectively. For the filtering experiment, the frequency range of 0.25 to 8 [kHz] is divided into $N = 12$ bands of equal length on the basilar membrane of the human cochlea. The consonant region of fricatives, characterized by a diffuse frication noise, can be spread across a wide frequency range. For perceptual cues covering $N > 1$ bands, it is assumed that each band contributes equally to the identification of target sound. For each individual utterance, let $P_c^{\mathrm{MAX}} = \max(P_c(\text{full-band}),\ P_c(f_1^H),\ \ldots,\ P_c(f_9^H),\ P_c(\mathrm{f}_1^L),\ \ldots,\ P_c(f_9^L))$, the maximum P_c over the FB, all nine high-pass, and all nine low-pass filtering conditions. According to Fletcher and Galt's product rule (Allen, 1994), the detection threshold for each band τ_b is related to a FB threshold of $\tau = 0.9P_c^{\mathrm{MAX}}$ by $(1 - 0.9P_c^{\mathrm{MAX}}) = (1 - \tau_b)^N$, solving for the detection threshold gives

$$\tau_b = 1 - \sqrt[N]{1 - 0.9P_c^{\mathrm{MAX}}}. \tag{1}$$

When there is no frequency range within which both $P_c(f^H)$ and $P_c(f^L)$ are greater than $0.9P_c^{\mathrm{MAX}}$, the cue region spans $N > 2$ bands, and a more conservative estimate of the frequency range is used. The non-cue regions provide a wider frequency range for the possible perceptual cue region. Bounds on the high- and low-frequency non-cue regions are determined by finding the frequencies such that $P_c < 3P_{\mathrm{chance}}$, $P_{\mathrm{chance}} = 1/16$, for the high- and low-pass data, respectively. Both cue region detection and rejection thresholds were chosen empirically.

Thus, the frequency boundaries of the spectral cue region $[f_{\mathrm{low}}, f_{\mathrm{high}}]$ can be estimated by

$$f_{\mathrm{low}} = \hat{f}^L : P_c(\hat{f}^L) \geq \tilde{\tau},$$

$$f_{\mathrm{high}} = \hat{f}^H : P_c(\hat{f}^H) \geq \tilde{\tau},$$

with $\tilde{\tau}$ defined as either τ_b or $3P_{\mathrm{chance}}$.

When the speech token has a low P_c even in the quiet, FB, and full-duration condition, a cue region cannot be isolated by the 3DDS method since the listeners will not show correct perception at any condition.

III. RESULTS

Next, we demonstrate how the perceptual cues of fricative consonants are isolated by the 3DDS method. For each

consonant, a single representative utterance (out of the six utterances) for each CV is presented in a figure and analyzed in detail. The results of experiments TR07, HL07, and MN05 are depicted as CPs; a CP displays the probability of hearing all possible responses r, given the spoken consonant s, as the conditions for a single experiment vary (Allen, 2005). More precisely, the CPs $P_{rls}(t)$, $P_{rls}(f)$, and P_{rls}(SNR) are shown for experiments TR07, HL07, and MN05, respectively, in Figs. 2–4. Confusions with probability <0.2 and Noise Only responses are not shown in the CPs in order to clearly display the primary confusions.

The figures are organized into three pairs /ʃ, ʒ/, /s, z/, and /f, v/ (Figs. 2–4) to highlight both the similarities and differences. The paired unvoiced–voiced fricatives are displayed in subfigures (a) and (b), respectively. Each of the subfigures contains five panels labeled with a boxed number in the upper left or right corner. Panel 1 shows the AI-gram of the full-bandwidth, full-duration CV at 18 [dB] SNR with the 3DDS-isolated spectral cue region highlighted by a small rectangular box; the range of truncation times is marked by a large frame. Panel 2, aligned with panel 1 along the time axes, shows the listener CP as a function of truncation time, a star and vertical dotted line marks the truncation time at which P_c drops below 90%, a second vertical dotted line marks the end of the frication region. The line marked with "a" in panel 2 shows the probability of responding "Vowel Only." Panel 3, rotated by 90° clockwise and aligned with panel 1 along the frequency axes, illustrates the CP when listeners hear a high- or low-pass filtered sound as a function of the nine cutoff frequencies, with dashed lines indicating high-pass responses and solid lines indicating low-pass responses. Panel 4 shows the CPs when the utterance is masked by white noise at six different SNRs, with the SNR_{90} marked by a star. The AI-grams, at each tested SNR, are displayed in panel 5. For the AI-grams of panel 5, only the region within the range of truncation times is shown.

A. /ʃɑ/ and /ʒɑ/

The results of the perceptual experiments for /ʃɑ/ and /ʒɑ/, for talker m118, are shown in Fig. 2.

The AI-gram for /ʃɑ/ from talker m118 (Fig. 2a.1) shows a wide-bandwidth, sustained frication noise above 2 [kHz] in the consonant region. The truncation experiment (TR07, Fig. 2a.2) shows that when the duration of the frication is truncated from the original ≈20 [cs] to ≈8 [cs], the listeners begin to show confusions with /ʒɑ/. Once the frication region is truncated to ≈1 [cs] (a burst), the listeners report /dɑ/ 80% of the time. The results from the filtering experiment (HL07, Fig. 2a.3) show that the perceptual cue region lies between 2 and 3.4 [kHz]. When the utterance is low-pass filtered below $f^L < 1$ [kHz], confusions with /fɑ/ emerge. High-pass filtering above $f^H > 3.9$ [kHz] causes confusions with /sɑ/. The low-pass filtering responses indicate that this energy above 3.9 [kHz] is unnecessary for correct perception, indicating that this high-frequency frication

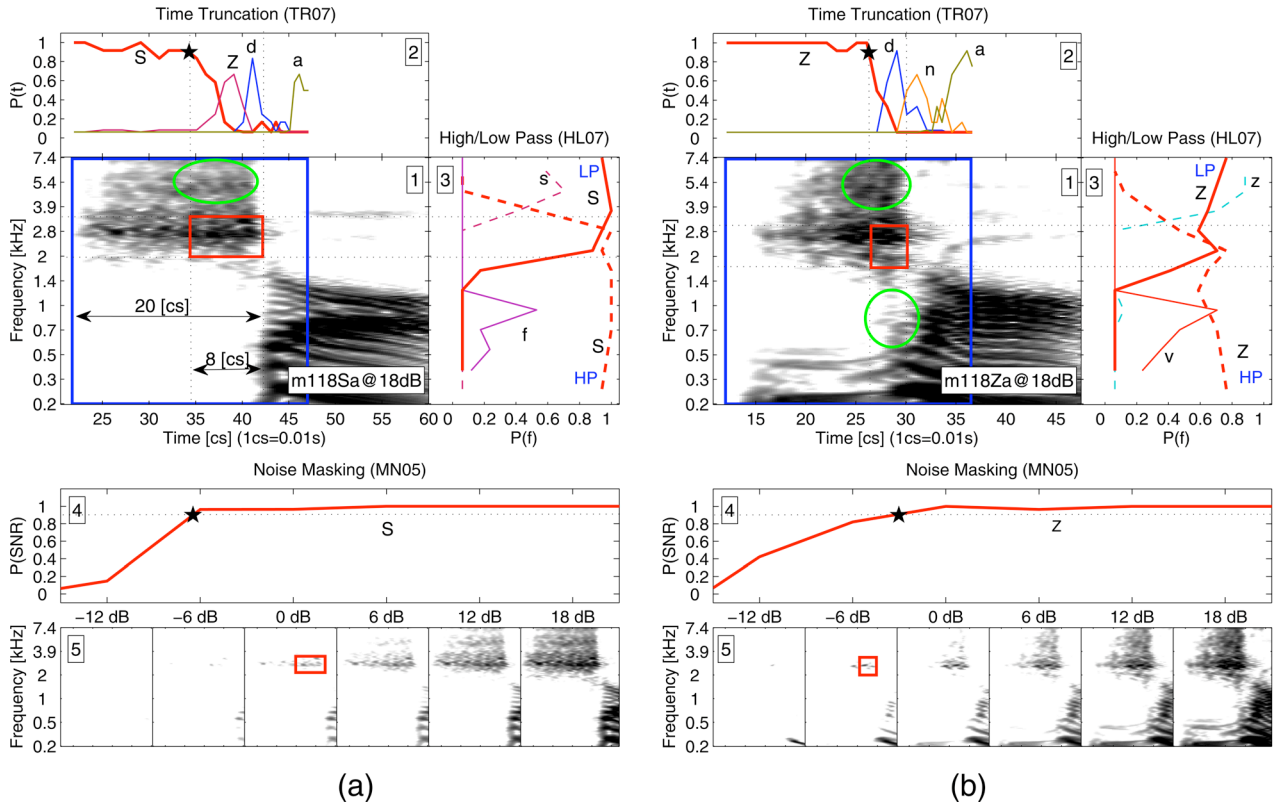

FIG. 2. (Color online) Isolated perceptual cue regions for /ʃɑ/ and /ʒɑ/, denoted by *S* and *Z*, respectively. Each utterance has five numbered panels: 1) the AI-gram with the highlighted perceptual cue region (rectangle), hypothetical conflicting cue region (ellipse), and the range of truncation times (large frame); 2) CP of TR07, aligned to panel 1 along the time axes, a represents the IPA vowel /ɑ/; 3) CP of HL07, rotated by 90° clockwise and aligned to panel 1 along the frequency axes; 4) CP of MN05; 5) AI-grams across the different tested SNRs (large frame region) showing the change of speech audibility as noise level increases.

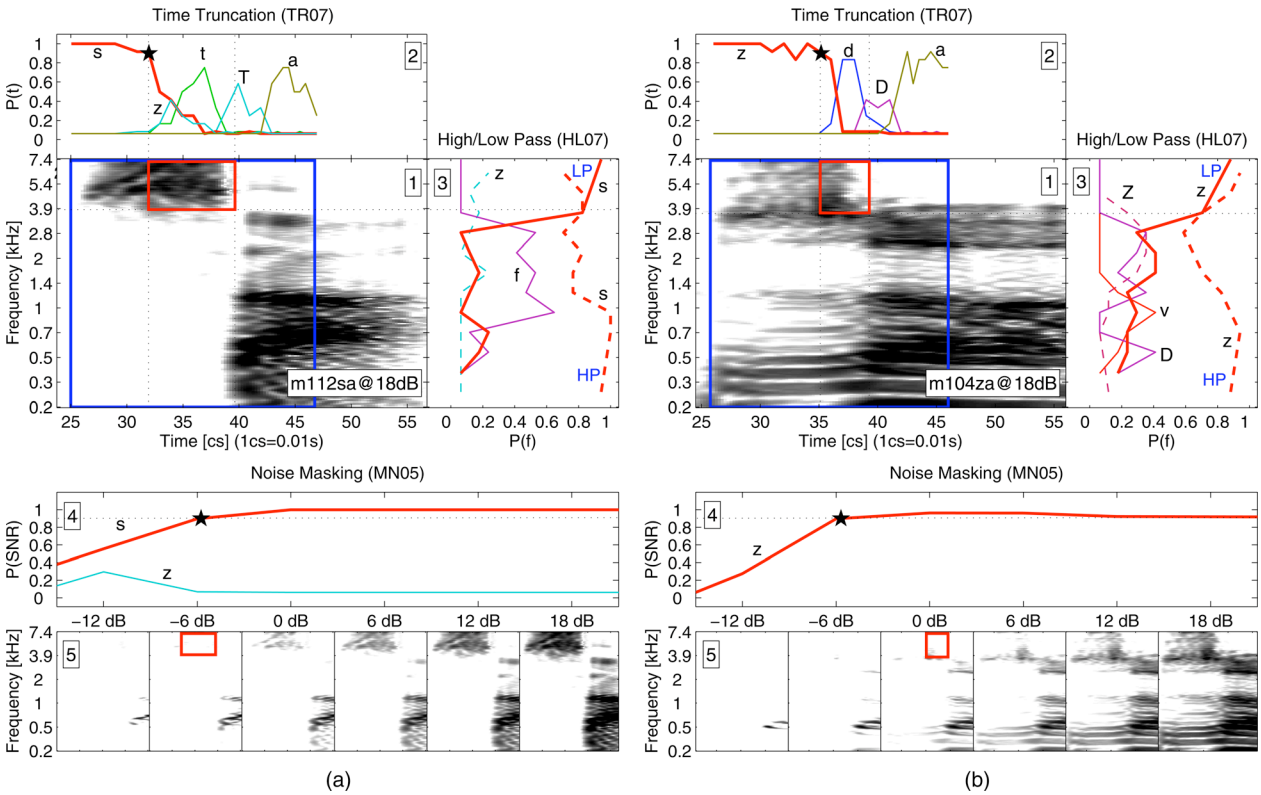

FIG. 3. (Color online) Isolated perceptual cue regions for /sɑ/ and /zɑ/. Each utterance has five numbered panels: 1) the AI-gram with the highlighted perceptual cue region (rectangle) and the range of truncation times (large frame); 2) CP of TR07, aligned to panel 1 along the time axes; 3) CP of HL07, rotated by 90° clockwise and aligned to panel 1 along the frequency axes; 4) CP of MN05; 5) AI-grams across the different tested SNRs (large frame region) showing the change of speech audibility as noise level increases. S, Z, T, D and a represent the IPA symbols /ʃ, ʒ, θ, ð, ɑ/, respectively.

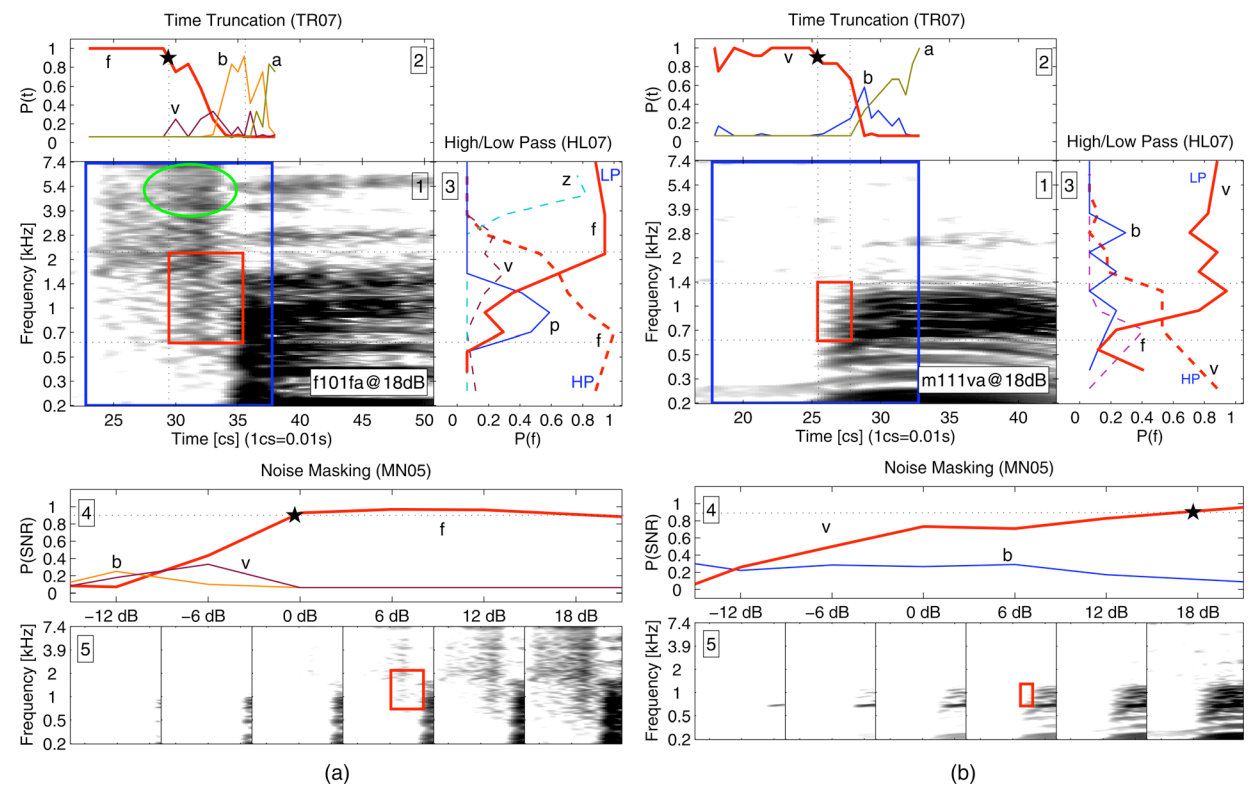

FIG. 4. (Color online) Isolated perceptual cue regions for /fɑ/ and /vɑ/. Each utterance has 5 numbered panels: 1) the AI-gram with the highlighted perceptual cue region (rectangle), hypothetical conflicting cue region (ellipse), and the range of truncation times (large frame); 2) CP of TR07, aligned to panel 1 along the time axes, a represents the IPA symbol /ɑ/; 3) CP of HL07, rotated by 90° clockwise and aligned to panel 1 along the frequency axes; 4) CP of MN05; 5) AI-grams across the different tested SNRs (large frame region) showing the change of speech audibility as noise level increases.

contains conflicting cues. The noise masking experiment (MN05, Fig. 2a.4) shows a sharp drop in the P_c, from 96% to 15% between −6 and −12 [dB] SNR. Examining the AI-gram across SNRs (Fig. 2a.5), we see a predicted loss of audibility of the isolated cue region between 0 and −6 [dB] SNR.

For the voiced /ʒɑ/ from talker m118, the AI-gram (Fig. 2b.1) for the consonant region contains a wide-bandwidth, sustained frication region above 1.6 [kHz] (similar to /ʃɑ/) along with a coincident voicing below 0.6 [kHz]. The truncation experiment (TR07, Fig. 2b.2) shows that the original ≈16 [cs] duration of the voiced frication can be shortened to ≈3.5 [cs] before listeners report non-target consonants. When the voiced frication is truncated to <2 [cs] listeners primarily report /dɑ/. Once the frication region is removed by truncation, leaving just the vowel onset, listeners primarily report /nɑ/ and do not report Vowel Only until the vowel onset is also removed. The results of filtering (HL07, Fig. 2b.3) show that the necessary perceptual cue region for /ʒɑ/ lies between 1.7 and 3 [kHz]. When the primary cue region is removed by a low-pass filter at $f^L \approx 1$ [kHz], listeners primarily report /vɑ/. When the utterance is filtered above $f^H > 3.2$ [kHz], listeners primarily report /zɑ/. The frication region above 3.2 [kHz] contains conflicting cues for /zɑ/, the P_c does not significantly change when this high-frequency frication noise is removed by low-pass filtering. The results of the noise masking experiment (MN05, Fig. 2b.4) show that the P_c begins to drop at 0 [dB] SNR. The AI-gram across SNRs (Fig. 2b.5) predicts full loss of audibility of the identified frication region below −6 [dB] SNR; correspondingly, the P_c falls from 82% at −6 [dB] SNR to 42% at −12 [dB] SNR.

The lower HL07 performance for /ʒɑ/ from talker m118 at the FB case ($P_c = 80\%$ in HL07 vs 100% at the corresponding MN05 and TR07 control conditions for different listener populations) is due to a percentage of listeners (3 out of 17) that had consistent difficulty discriminating /ʒɑ/ across all of the utterances, despite little to no error in the FB perception of other high to medium scoring fricatives in our data set. For these listeners, since /ʒ/ does not have a dedicated representation in written English (e.g., "sh" for /ʃ/), it seems likely that these low scores are the result of an insufficient practice session.

It may be surprising to see that perception of the target sound /ʒɑ/ does not become confused with /ʃɑ/ when the low-frequency (<0.7 [kHz]) voicing is removed by a high-pass filter (Fig. 2b.3). Further analysis revealed that the frication regions of voiced sibilants are amplitude modulated by $F0$, the fundamental frequency of vocal vibration, retaining a sufficient amount of voicing information for correct perception even when the low-frequency voicing is removed (Fig. 2b.3). A detailed summary on the perceptual cues for discrimination of voicing can be found in Sec. IV B.

1. Other utterances

3DDS analysis of the five other /ʃɑ/ utterances show similar cue region results to those for the utterance from talker m118. The truncation data show that the P_c drops below 90% once the consonant frication regions are truncated to a duration of <9 [cs], on average. The frequency range of the cue region was found to be between 1.6 and 4.2 [kHz] for the total group of utterances, with variability within that range across different talkers. All six /ʃɑ/ utterances resulted in a $P_c > 90\%$ for the noise masking experiment at 12 [dB] SNR, with a perceptual robustness to noise that varied from −7 to 5 [dB], as quantified by SNR_{90}.

Of the remaining five /ʒɑ/ utterances, four (talkers f103, m107, m114, and m117) show cue region results similar to the utterance from talker m118. The truncation results show that the consonant is perceived correctly until the frication region is shortened to <3.5 [cs], on average. The overall frequency range of the cue region lies between 1.4 and 3.9 [kHz], with variability within this range across talkers. The noise masking experiment shows that the robustness to noise of the perceptual cues, as measured by SNR_{90}, falls in a 12 [dB] range, from −7 to 5 [dB] SNR. One /ʒɑ/ utterance (talker f108) has a P_c of only 40% at the quiet, FB, full-duration condition. Analysis of the AI-gram of this utterance revealed a frication with similar frequency distribution and duration to the other /ʒɑ/ utterances, but with a barely audible voicing (as estimated by the AI-gram), predicting a high likelihood of confusion with /ʃɑ/; these voicing confusions are observed in the listener responses (50% /ʃɑ/ in quiet).

Summary cue region results are provided in Table I.

B. /sɑ/ and /zɑ/

The results of the perceptual experiments for /sɑ/ (talker m112) and /zɑ/ (talker m104) are shown in Fig. 3.

TABLE I. Summary of 3DDS results for /ʃɑ, ʒɑ, sɑ, zɑ/. Minimum duration defined from the truncation time at which perception drops below 90% to end of frication. Utterances ordered based on SNR_{90} value. Only utterances with conclusive 3DDS estimates from all three experiments are listed.

CV	Talker	Min Dur [cs]	Freq [kHz]	SNR_{90}
/ʃɑ/	m118	8	2.0–3.4	−7
	m115	9	1.6–3.6	−2
	m111	8.5	2.0–4.2	−1
	f103	8	2.0–3.7	−1
	f109	11	2.5–3.6	0
	f106	9	1.9–4.2	5
/ʒɑ/	f103	4	1.9–3.7	−7
	m118	3.5	1.7–3.0	−3
	m114	3	1.4–3.1	−1
	m117	3	2.6–3.9	3
	m107	2.5	2.1–3.7	5
/sɑ/	f109	7.5	6.0–8.0	−10
	m112	7.5	3.9–8.0	−6
	f113	10.5	5.4–8.0	−5
	f108	9	5.4–8.0	−2
/zɑ/	m120	5	3.9–8.0	−10
	f105	4	5.4–8.0	−7
	m104	4	3.6–8.0	−6
	f108	4.5	3.8–8.0	−1
	m118	5.5	4.6–8.0	−1

The AI-gram for /sɑ/ from talker m112 (Fig. 3a.1) displays a high-frequency, sustained frication noise above 3.2 [kHz] in the consonant region. The truncation results (Fig. 3a.2) show that once the duration of the frication is truncated from the original ≈14 [cs] to < 7.5 [cs], the P_c begins to drop, with a high percentage of listeners reporting /tɑ/ once the frication is shortened below 4 [cs]. The results of the filtering experiment (Fig. 3a.3) show that the cue region lies above 3.9 [kHz]. Low-pass filtering of the utterance at $0.9 \leq f^L \leq 2.8$ causes confusion with /fɑ/. The noise masking experiment (Fig. 3a.4) shows a drop in the P_c after 0 [dB] SNR, with an SNR_{90} of approximately −6 [dB] SNR. This is consistent with the AI-grams across SNRs (Fig. 3a.5), which predict a faint but still audible frication within the 3DDS-isolated cue region at −6 [dB] SNR.

For the voiced /zɑ/ from talker m104, the AI-gram (Fig. 3b.1) displays a sustained frication region above 2.3 [kHz] and coincident voicing mainly below 0.7 [kHz], in the consonant region. The time truncation results (Fig. 3b.2) show that once the duration of the frication is truncated from the original ≈14.5 [cs] to ≤4 [cs], listeners begin to report non-target consonants. Once the frication is truncated to ≤3 [cs], listeners primarily report /dɑ/. The results of the filtering experiment (Fig. 3b.3) show that the cue region for this /zɑ/ utterance falls between 3.6 and 8 [kHz]. No strong confusion emerges when the spectral cue region is removed by filtering, instead the listeners chose the Noise Only response. The noise masking experiment (Fig. 3b.4) shows an abrupt drop in P_c below −6 [dB] SNR. The AI-grams across SNRs (Fig. 3b.5) predict that a small amount of the isolated cue region is still audible at 0 [dB] SNR and is completely masked by noise at −12 [dB] SNR.

1. Other utterances

Of the remaining five /sɑ/ utterances, three of them (talkers f108, f109, and f113) have similar cue regions as the utterance from talker m112. The time truncation data for these utterances showed a drop in the P_c once the frication region was truncated below 9 ± 1.5 [cs]. The frequency ranges of the perceptual cue regions are between 3.9 and 8 [kHz], with utterance-specific variability within this region. The SNR_{90} measurements fall across an 8 [dB] range, from −10 to −2 [dB] SNR. The remaining two /sɑ/ utterances, selected for their low perceptual scores, were primarily confused with /zɑ/ in quiet. Further analysis showed barely audible voicing cues in the utterance from talker m111 and a low-level, short-duration (4 [cs]) frication from talker m117.

Of the five remaining /zɑ/ utterances, four (talkers f105, f108, m118, and m120) contain similar perceptual cue regions to those of talker m104. The truncation data for these utterances showed a sharp drop in P_c when the frication was shortened to 5 ± 1 [cs] and the frequency ranges of the cue regions fell between 3.6 and 8 [kHz]. The noise masking experiment resulted in SNR_{90} measurements within a 9 [dB] range, from −10 to −1 [dB] SNR. One utterance (talker f109) showed strong (≈50%) confusion with /ðɑ/ in quiet, yet P_c rose to 100% once the first quarter of the consonant region was removed by truncation. Further investigation showed an energy burst before the onset of frication, creating a conflicting cue region, which led to this /ðɑ/ confusion. This conflicting cue region led to inconclusive filtering and noise-masking results for this utterance.

Summary cue region results are provided in Table I.

C. /fɑ/ and /vɑ/

The results of the perceptual experiments for /fɑ/ (talker f101) and /vɑ/ (talker m111) are shown in Fig. 4.

The AI-gram for /fɑ/ from talker f101 (Fig. 4a.1) displays a wide-bandwidth, sustained frication noise that spans 0.3 to 7.4 [kHz], in the consonant region. The truncation results (Fig. 4a.2) show that once the frication is shortened from the original ≈12 [cs] to <6.5 [cs], listeners report non-target consonants. Truncating the entire frication region while leaving the vowel onset intact, results in a large proportion of /bɑ/ responses (>80%). The filtering experiment (Fig. 4a.3) shows that the frequency range of the cue region, despite the wide-bandwidth of the full frication region, lies between 0.6 and 2.2 [kHz]. High-pass filtering at $f^H \geq 3.9$ [kHz] causes listeners to primarily report /zɑ/, indicating that the frication energy above this frequency contains a conflicting cue. Low-pass filtering at $0.7 \leq f^L \leq 1.3$ [kHz] results in listeners reporting /pɑ/, indicating that the full wide bandwidth of the cue region is necessary for perception of this /fɑ/ utterance. The noise masking experiment (Fig. 4a.4) shows an SNR_{90} of −1 [dB]. The AI-grams across SNRs (Fig. 4a.5) predict a loss of audibility between 6 and 0 [dB] SNR.

For the voiced /vɑ/ from talker m111, the AI-gram (Fig. 4b.1) displays a faint wide-band frication and a coincident low-frequency (<0.3 [kHz]) voicing in the consonant region. The vertical dotted line marking the end of the frication region (at ≈27 [cs]) indicates that the frication is briefly sustained into the onset of the vowel (as determined from the time waveform). The time truncation results (Fig. 4b.2) show that once the frication region is shortened from the original ≈11 [cs] to <2 [cs], listeners report confusions with /bɑ/. The results of the filtering experiment (Fig. 4b.3) show that the perceptual cue region lies between 0.6 and 1.4 [kHz]. Filtering out the isolated cue region leads to Noise Only responses. The noise masking experiment (Fig. 4b.4) suggests that the perceptual cue for this utterance can be masked by low levels (12 [dB] SNR) of white noise. The AI-gram across SNRs (Fig. 4b.5) predicts that some of the isolated cue remains audible up to 6 [dB] SNR but is shortened in duration to <2 [cs]. In agreement with the results of the truncation experiment, the primary confusion at this 6 [dB] SNR condition is with /bɑ/.

1. Other utterances

The /fɑ/ utterance from talker m111 contains a similar spectral cue region to that of talker f101. A wide-band frication in the slightly lower frequency range of 0.7 to 2 [kHz], a minimum frication duration for correct recognition of 4.5 [cs], and an SNR_{90} of 10 [dB] summarize the 3DDS findings

for the cue region of this utterance. The low-performance /fɑ/ utterance from talker m117 showed a $P_c < 90\%$ correct even in the quiet, unmodified condition, leading to inconclusive 3DDS results. The three remaining low and medium-performance /fɑ/ utterances (talkers f103, f105, and m112) have low-level but audible frication regions. These low-level but audible regions were erroneously removed in the preparation of the waveforms for the experiments, leading to a $P_c < 90\%$ at the control condition. As a result, we were unable to isolate the perceptual cues for these three /fɑ/ utterances. We only present results unaffected by this inappropriate consonant signal processing.

The /vɑ/ utterance from talker f108 is defined by a spectral cue region that is almost the same as the one observed for talker m111. This utterance has a minimum frication duration of 1.5 [cs] for correct perception, a frequency range of 0.5 to 1.1 [kHz], and an SNR_{90} of 3 [dB] SNR. Of the remaining /vɑ/ utterances, three (talkers f105, m104, and m120) are composed of consonant cues that are partially masked by the 12 [dB] of noise used in the control condition, leading to inconclusive 3DDS results. One mislabeled /vɑ/ utterance in the data set (talker f103) is primarily reported as a /fɑ/ in quiet and low noise levels.

Summary cue region results are provided in Table II.

D. /θɑ/ and /ðɑ/

For all utterances of both the voiceless dental /θɑ/ and voiced dental /ðɑ/, significant confusions were seen at the control conditions for all three experiments. The maximum P_c of any single utterance at the control condition (12 [dB] SNR, FB, no truncation) for any experiment was 80%, with wide variability of baseline performance across the other experiments; thus, no spectral region could be isolated that contained the necessary cues for perception. The low recognition scores can be attributed to many factors that vary for each utterance including low-level frication, ambiguous voicing, mislabeling of utterances (experimental and practice sets), and listener difficulty in discriminating the two consonants. The listener difficulty can be due either to a lack of perceptual awareness as to the differences between the two consonants or to an insufficient practice session.

Due to the low perceptual scores at all experiment conditions, a region of perceptual cues cannot be isolated with the 3DDS method. Future investigations on the perceptual cues of /θɑ, ðɑ/ will need to carefully control for a data set that can be reliably recognized by American English listeners.

TABLE II. Summary of 3DDS results for /fɑ,vɑ/. Minimum duration defined from the truncation time at which perception drops below 90% to end of frication.

CV	Talker	Min Dur [cs]	Freq [kHz]	SNR_{90}
/fɑ/	f101	6.5	0.6–2.2	−1
	m111	4.5	0.7–2.0	10
/vɑ/	f108	1.5	0.5–1.1	3
	m111	2	0.6–1.4	18

E. Consistency and variability

Using the duration, frequency range, and robustness to noise estimates from each experiment, we next examine the variability of the 3DDS-isolated perceptual cues across speakers.

The sibilant fricatives /ʃ, ʒ, s, z/ each show consistency across speakers in terms of minimum duration for correct perception and frequency range of the cue region as seen in Table I. For /ʃɑ/, the cue regions of all utterances fell within the frequency range of 1.6 to 4.2 [kHz] and the minimum frication durations were measured to be 9.5 ± 1.5 [cs]. Five /ʒɑ/ utterances contained necessary cues within the frequency range of 1.4 to 3.9 [kHz], with minimum durations of 3.5 ± 1 [cs]. The four robust /sɑ/ utterances contained necessary cues which fell within the frequency range of 3.9 to 8 [kHz] with a minimum duration of 9 ± 1.5 [cs]. For /zɑ/, the five robust utterances contained necessary cues in the frequency range of 3.6 to 8 [kHz] with a minimum duration of 5 ± 1 [cs]. An analysis of variance (ANOVA) of the center frequency values calculated from the filtering experiment shows significant differences across the sibilant consonants $F(3,16) = 161.87$, $p < 0.001$. A multiple comparison test shows significant differences in the frequency cues for two groups, the cue regions for /ʃɑ, ʒɑ/ are significantly lower in frequency than the isolated cue regions for /sɑ, zɑ/. Thus, similar necessary frequency cues are shared by /ʃɑ, ʒɑ/ and by /sɑ, zɑ/. In terms of duration cues, the voiced fricatives /ʒɑ, zɑ/ can withstand more truncation (2.5 to 5.5 [cs]) than their unvoiced counterparts /ʃɑ, sɑ/ (7.5 to 11 [cs]) before recognition is affected. An ANOVA of the estimates of minimum duration showed significant differences across the sibilant consonants $F(3,16) = 43.36$, $p < 0.001$. A multiple comparison test shows a significant difference between the minimum duration of two groups, the unvoiced /ʃɑ, sɑ/, and the voiced /ʒɑ, zɑ/. An ANOVA of the SNR_{90} values for the sibilant fricatives did not show any significant differences.

The variability of the non-sibilants /fɑ/ and /vɑ/ is more difficult to quantify as there were only two utterances per CV with sufficient perceptual cues. Table II provides the details of the perceptual cue regions for these utterances. The spectral cue region estimates for /fɑ/ are relatively similar, differing by only 2 [cs] for minimum duration and overlapping over the majority of the frequency ranges. For /vɑ/, the minimum duration estimates differed by 0.5 [cs] and the frequency estimates overlap. An ANOVA of the results summarized in Table II showed no significant difference between the unvoiced /fɑ/ and voiced /vɑ/ in terms of minimum duration or SNR_{90}. The ANOVA did show a significant difference $F(1,2) = 21.24$, $p = 0.044$ for the center frequencies of /fɑ/ and /vɑ/. Due to the limited amount of data for the non-sibilants, the full variability of these cue regions is not considered to have been fully explored.

F. Robustness to noise

What is the underlying physical (i.e., acoustical) basis for the noise-robustness of fricative consonants? The SNR_{90} is used to quantify the noise robustness of an utterance; it is

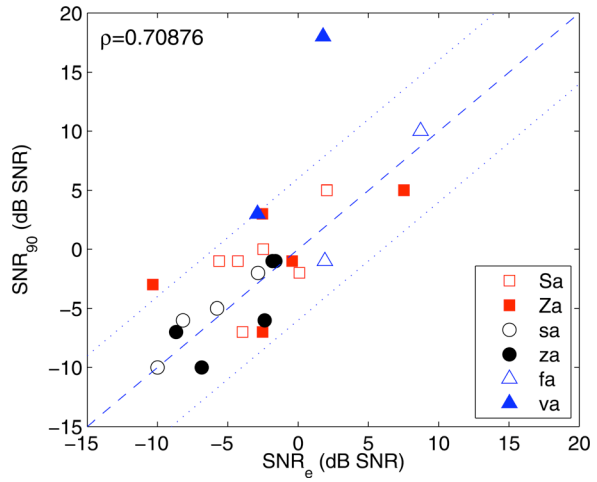

FIG. 5. (Color online) Pearson correlation between SNR_e and SNR_{90} $\rho = 0.71$ (p-value $= 1 \times 10^{-4}$). All utterances with 3DDS-isolated cue regions are reported. S, Z, and a are used to represent the IPA symbols /ʃ, ʒ, ɑ/, respectively.

defined as the SNR above which listeners can correctly perceive the sound with a $P_c > 90\%$. For a particular consonant, the SNR_{90}'s can exhibit significant variation across different talkers. Qualitative analysis of the noise masking data from experiment MN05 and the associated AI-grams at various SNRs (panels 4 and 5 in Figs. 2–5) show that fricatives with lower perceptual SNR_{90} values (more robust to noise) have 3DDS-isolated cue regions that are higher in intensity.

Previous studies (Régnier and Allen, 2008; Li *et al.*, 2010) found that the perceptual SNR_{90} for stop consonants was correlated to a physical measure, the predicted audible threshold (SNR_e) of the 3DDS-isolated cue region. A similar relationship was found to also apply to fricative consonants. Figure 5 shows a scatter plot of SNR_{90} vs SNR_e for six of the eight fricatives. SNR_e is estimated by calculating the lowest SNR at which audibility of the 3DDS-isolated cue region is predicted by the AI-gram. Audibility for this measure is averaged over a critical band and 5 [cs] in time, similar to Régnier and Allen (2008). The utterances with recognition scores below 90% at the FB, full-duration, 12 [dB] SNR condition do not have 3DDS-isolated cue regions to report.

The Pearson correlation between the perceptual SNR_{90} and physical SNR_e is $\rho \approx 0.71$. As seen in Fig. 6, SNR_e is significantly correlated with SNR_{90}. Thus, the audibility of the 3DDS-isolated cue region can predict the perceptual robustness of a consonant to masking noise, suggesting that audibility of the cues contained in the 3DDS-isolated spectral region is necessary for correct perception. As shown by the dotted lines, a ±6 [dB] difference exists between estimates of the two thresholds.

IV. DISCUSSION

In this study, we generalized the 3DDS psychoacoustic method to fricative American English consonants. This method allows us to examine the effects of highly variable natural speech components on human perception. We have also identified several natural confusions that are observed for these fricatives under different modification conditions.

A. Discriminating cues for the place of articulation

For all of the fricatives in this study, the 3DDS-isolated cue regions are within the frication region. This is consistent with the observations of Harris (1958) for /s, ʃ, z, ʒ/+/i, e, o, u/. The alveolar consonants /sɑ, zɑ/ have isolated cue regions in the sustained frication, no lower than 3.6 [kHz]. The palato-alveolar consonants /ʃɑ, ʒɑ/ have isolated cue regions in the sustained frication between 1.4 and 4.2 [kHz]. For the non-sibilant labiodentals /fɑ, vɑ/, a band of frication between the frequency range of 0.5 and 2.2 [kHz] is isolated as the cue region. Frication noise at higher frequencies than the isolated cue regions is present in all utterances of /ʃɑ, ʒɑ, fɑ, vɑ/, but does not help or hinder correct perception. Thus, a necessary perceptual cue for fricative place of articulation is the frequency of the lowest bound (i.e., the frequency of the lower edge) of the band of frication noise.

B. Discriminating cues for the manner of articulation

For all utterances in this study, the minimum frication durations were identified, as summarized in Tables I and II. When the frication is truncated beyond these minimum durations, but not completely removed, strong (>50%) confusions with plosives emerge. The plosive confusions for severely

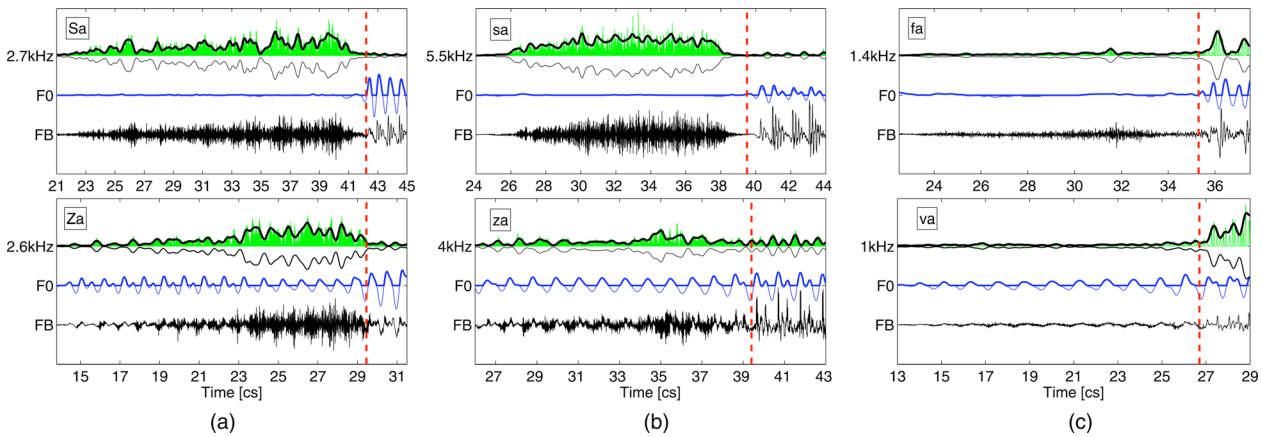

FIG. 6. (Color online) The FB signal. $F0$ (low-pass filtered, $f^L_{\text{cutoff}} = 0.5$ [kHz]), and the band-pass filtered frication noise (f_{center} marked on the y-axis), for unvoiced /ʃɑ, sɑ, fɑ/ and voiced /ʒɑ, zɑ, vɑ/. Vertical dashed line marks the vowel onset. S, Z, and a are used to represent the IPA symbols /ʃ, ʒ, ɑ/, respectively.

truncated /zɑ, ʒɑ/ utterances are /dɑ, gɑ/. For /sɑ, ʃɑ/ the confusions are /dɑ, gɑ, tɑ/. For both /fɑ, vɑ/, only strong /bɑ/ confusions are observed when the frication is truncated beyond the minimum duration. Thus, when the frication is truncated to <3 [cs], the spectral information that remains is perceived primarily as a voiced plosive. The durational cue, encoded in the sustained frication region, is a necessary cue for identification of fricatives.

C. Discrimination of voicing

For stop consonants, it is evident that timing information, such as voice-onset time, defined as the duration between the release of burst and the onset of voicing, is critical for the discrimination of voiced consonants /b, d, g/ from their unvoiced counterparts /p, t, k/ (Liberman *et al.*, 1958; Li and Allen, 2011). Three characteristics are observed in the consonant region of naturally-produced voiced fricatives: *F*0 modulation of the frication noise, coincident low-frequency voicing, and a longer original duration than unvoiced homorganic fricatives. The results of the 3DDS psychoacoustic experiments provide information about the relative perceptual roles of these three characteristics.

In Sec. III, we observed that unvoiced sibilants /ʃɑ, sɑ/ tend to have a longer frication region than their voiced counterparts /ʒɑ, zɑ/. Most studies note that the duration of unvoiced fricatives is generally longer than that of the voiced fricatives (Baum and Blumstein, 1987; Stevens *et al.*, 1992; Jongman *et al.*, 2000) but the natural distributions of the two categories overlap considerably (Baum and Blumstein, 1987). The results of the truncation experiment (TR07) show that even when the unvoiced fricatives are truncated to the shorter average original duration of voiced fricatives, listeners can still correctly perceive the unvoiced fricatives, demonstrating that duration in natural speech is not a necessary cue for voicing discrimination. Only when the unvoiced /ʃɑ, sɑ/ are deeply time-truncated to a duration similar to the minimum possible duration of /ʒɑ, zɑ/ (2.5 to 5.5 [cs]), do listeners report some confusions with the corresponding voiced fricative. Similarly, when the frication for the unvoiced /fɑ/ is truncated to ≤3 [cs], but not completely removed, weak (<40%) confusions with the voiced fricative /vɑ/ are reported. Together these observations suggest that, although duration may play a role when the signal is sufficiently degraded, it is not the necessary discriminating cue for fricative voicing.

In the filtering experiment (HL07), the P_c for all voiced fricative /zɑ/ utterances with isolated cue regions remains >80% even when the original signal is high-pass filtered at 1.3 [kHz]. Similarly, the P_c of all /ʒɑ/ utterances does not drop below the FB level until the signal is high-pass filtered above 0.9 [kHz]. These results are evidence that the low-frequency voicing is also not a necessary cue. Thus, the *F*0 modulation of the remaining consonant region, the frication noise, is the cue that is necessary for correct perception of voicing.

The modulation of the frication noise of the voiced fricatives /ʒɑ, zɑ, vɑ/, and the lack of modulations for the unvoiced fricatives /ʃɑ, sɑ, fɑ/, can be observed in Fig. 6. Figure 6 displays plots of the FB signal, the fundamental (*F*0), and the band-pass filtered signal of the frication along with the envelope, for the six fricatives (from the same talkers as Figs. 2–4). Previous studies have shown that the modulations in speech are perceptually significant; a modulation filter bank model would provide a more detailed representation of how such modulations are perceived in noise (Jørgensen and Dau, 2011).

D. Conflicting cue regions

The masking of speech by noise, and the resulting confusions, are of key importance to understanding speech communication. In a previous study (Li *et al.*, 2010), it was discovered that naturally-produced stop consonants often contain acoustic components that are not necessary for correct perception but can cause listeners to confuse the target sound with competing sounds once the primary cue region for the target consonant is removed. A speech component that contains cues for a non-target consonant and is also not necessary for correct perception of the target consonant is denoted as a *conflicting cue region*. For instance, in Fig. 4(a) of Li *et al.* (2010), the /kɑ/ utterance from talker f103, with a mid-frequency burst centered at 1.6 [kHz] isolated as the cue region, also contains a high-frequency burst energy above 4 [kHz] and a low-frequency burst energy below 1 [kHz] that contain perceptual cues for /tɑ/ and /pɑ/, respectively. Once the mid-frequency burst cue region is removed, /kɑ/ is confused with /tɑ/ or /pɑ/ (Li and Allen, 2011); selective amplification of the conflicting cue regions can lead to complete morphing of the utterance into a consistently-perceived /tɑ/ or /pɑ/ (Kapoor and Allen, 2012).

Fricative consonants can also contain conflicting cue regions, specifically, all /ʃɑ/ and most of the /ʒɑ/ utterances that we examined contain conflicting cue regions in the frication above 4 [kHz]. When the frication <4 [kHz] is removed by filtering from /ʃɑ, ʒɑ/, the listeners report the non-target consonants /sɑ, zɑ/, respectively. The non-sibilant fricative /fɑ/ from talker f101 [Fig. 4(a)] also contains a high-frequency frication noise above 3 [kHz] that leads most listeners to report /zɑ/ in the absence of the mid-frequency /fɑ/ cue region. Similarly, the /fɑ/ utterance from talker m111 contains a high-frequency conflicting cue region for /zɑ/.

These conflicting cue regions often have a significant impact on speech perception, specifically when the primary cue region is masked under noisy circumstances. Based on our perceptual data, we hypothesize that many of the most frequent confusions (Miller and Nicely, 1955; Phatak *et al.*, 2008), e.g., /p/ ⟺ /t/ ⟺ /k/, /b/ ⟺ /d/ ⟺ /g/, /ʃ/ ⟺ /s/, /ʒ/ ⟺ /z/, /f/ ⟺ /v/ ⟺ /b/, and /m/ ⟺ /n/, are explained by the existence of such conflicting cues. Thus, the most efficient way to reduce confusion in speech perception is to either increase the strength of the primary spectral cue region and/or remove the conflicting cue region(s).

E. 3DDS method for isolating the perceptual cue region

The 3DDS method has proven to be effective in locating the spectral regions that contain the necessary cues for

correct perception along with conflicting cues for both plosives (Li *et al.*, 2010) and fricatives in natural speech. No single cue was found to be sufficient for the perception of a fricative, instead the combination of all necessary cues are together sufficient for perception. The 3DDS method allows one to analyze the perceptual effects of speech components without making assumptions about the time-frequency location or type of perceptual cues.

The 3DDS method finds the spectral region that contains all necessary cues. This region can contain multiple acoustic elements and variable cues (see Tables I and II). In this study, a sufficient set of necessary cues were isolated in a subset of the frication region of each utterance. Once the spectral region is isolated, further analysis (as in Secs. IV A, IV B, and IV C) is needed in order to determine the discriminating cues. Necessary cues that are identified in this study have been found to be encoded in the neural responses of auditory cortex (Mesgarani *et al.*, 2008).

Despite its success, the 3DDS method has limitations. We conclude that a requirement of 3DDS is that all test utterances be perceived correctly at the control condition of the three experiments. Utterances that do not meet this requirement have 3DDS results where no spectral cue region contains the full set of necessary cues. Our initial assumptions about the distribution of error, which led to the sampling of 1/3 high, medium, and low-scoring utterances for experiments TR07 and HL07, were therefore flawed. A recent study has shown that the majority of CV utterances are high-scoring for normal-hearing listeners in quiet conditions (Singh and Allen, 2012). Thus, the inclusion of 2/3 medium and low-scoring utterances was not representative of the general distribution.

A second weakness of this analysis is the audibility predictions based on the AI-gram, which could be improved; since the AI-gram is a linear model, it does not account for cochlear compression, forward masking, upward masking, and other cochlear and neural non-linear responses. Incorporation of these non-linear features will require considerable work. The current linear AI-gram model has proved useful in predicting audibility (Lobdell, 2009; Lobdell *et al.*, 2011). A key assumption of this study is that the normal hearing listeners are using the same cues for perception. It is possible that children and people with hearing loss may rely on different or additional cues for consonant perception.

A number of extensions to this work need to be carried out. Thus far, we have not successfully isolated perceptual cues for /θ, ð/. An extension of 3DDS to the nasals would also add to the current literature of speech perception. In addition to the vowel /ɑ/, the 3DDS method may be applied to other vowels. Analysis of conflicting cue regions can also be applied to investigations of noise-robustness and natural confusions. The perceptual differences between groups of speakers or listeners (e.g., English as a second language) could be assessed using the 3DDS method. Application of this method to different speaker types (e.g., clear vs normal speech) would allow one to see how the primary perceptual cue regions, and conflicting cue regions, are naturally modified by the speaker to specifically affect perception. Finally, although not the purpose of this study, a larger database of speech sounds would be necessary in order to further investigate the variability of perceptual cues across talkers.

V. SUMMARY AND CONCLUSIONS

In this study, the 3DDS method (Li *et al.*, 2010) is applied to the fricative consonants, extending a novel technique for isolating the cue regions of natural utterances. Both target-consonant cue regions and conflicting cue regions are analyzed. Conflicting cue regions in natural speech frequently correspond with the consonant confusions that arise under noisy or limited bandwidth conditions. The perceptual SNR_{90}'s are found to significantly correlate with the physical SNR_e's (derived from the AI-gram audibility predictions of the 3DDS-isolated cue region); this confirms that audibility of the cues in the 3DDS-isolated region is necessary for perception.

The perceptual roles of the many cues in naturally-produced fricative consonants are examined. Past studies have observed that the spectra of fricatives vary with place of articulation. The results of this study show that a subset of the frication region, the lowest frequency edge of the band of frication noise is the perceptually-relevant cue region. The durational cues are found to discriminate the class of fricatives from plosives with similar spectral characteristics. Past studies have observed that voiced fricatives exhibit modulations and are, in general, longer in duration than voiced homorganic fricatives. In our study, consonant duration is not the robust perceptual voicing cue; when voiceless fricatives are truncated to the same original durations as their voiced counterparts, the error remains at zero. The modulation of the frication noise provides the necessary cue for voicing; listeners correctly perceive the target consonant even when the entire low-frequency spectral region is removed by filtering but the modulations of the frication noise remain. It is concluded that the lower frequency bound of the frication noise, duration, and the modulation of the consonant region (for voiced fricatives) are all necessary cues for fricative perception.

Although natural speech signals can be highly variable, the consistency of cues within the 3DDS-isolated spectral regions provides clues about the workings of human speech recognition. Perceptually-based methods that do not make initial assumptions about the nature of perceptual cues, such as the 3DDS method, will help to answer the considerable amount of questions that remain in the field of speech perception.

ACKNOWLEDGMENTS

We would also like to thank the members of the HSR research group, particularly Roger Serwy, for their critical discussions. We acknowledge the support of Etymotic Research, Phonak, and the Linguistic Data Consortium for providing the database. This research was supported by Grant No. R21-RDC009277A from the National Institute of Health and by a grant from Research in Motion (RIM).

Allen, J. B. (**1994**). "How do humans process and recognize speech?" IEEE Trans. Speech Audio Process. **2**(4), 567–577.

Allen, J. B. (**2005**). "Consonant recognition and the articulation index," J. Acoust. Soc. Am. **117**(4), 2212–2223.
ANSI S3.5 (**1969**). *American National Standard Methods for the Calculation of the Articulation Index* (American National Standards Institute, New York).
Baum, S. R., and Blumstein, S. E. (**1987**). "Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English," J. Acoust. Soc. Am. **82**(3), 1074–1077.
Behrens, S., and Blumstein, S. E. (**1988**). "On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants," J. Acoust. Soc. Am. **84**(3), 861–867.
Bell, C. G., Fujisaki, H., Heinz, J. M., Stevens, K. N., and House, A. S. (**1961**). "Reduction of speech spectra by analysis-by-synthesis technique," J. Acoust. Soc. Am. **33**(12), 1725–1736.
Blumstein, S. E., Stevens, K. N., and Nigro, G. N. (**1977**). "Property detectors for bursts and transitions in speech perceptions," J. Acoust. Soc. Am. **61**(5), 1301–1313.
Cooper, F., Delattre, P., Liberman, A., Borst, J., and Gerstman, L. (**1952**). "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am. **24**(6), 597–606.
Delattre, P., Liberman, A., and Cooper, F. (**1955**). "Acoustic loci and translational cues for consonants," J. Acoust. Soc. Am. **27**(4), 769–773.
Fletcher, H., and Galt, R. (**1950**). "Perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**(2), 89–151.
French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**(1), 90–119.
Furui, S. (**1986**). "On the role of spectral transition for speech perception," J. Acoust. Soc. Am. **80**(4), 1016–1025.
Harris, K. S. (**1958**). "Cues for the discrimination of American English fricatives in spoken syllables," Lang Speech **1**(1), 1–7.
Hedrick, M. S., and Ohde, R. N. (**1993**). "Effect of relative amplitude of frication on perception of place of articulation," J. Acoust. Soc. Am. **94**(4), 2005–2026.
Heinz, J., and Stevens, K. (**1961**). "On the properties of voiceless fricative consonants," J. Acoust. Soc. Am. **33**(5), 589–596.
Hughes, G. W., and Halle, M. (**1956**). "Spectral properties of fricative consonants," J. Acoust. Soc. Am. **28**(2), 303–310.
Jongman, A. (**1988**). "Duration of frication noise required for identification of English fricatives," J. Acoust. Soc. Am. **85**(4), 1718–1725.
Kapoor, A., and Allen, J. B. (**2012**). "Perceptual effects of plosive feature modification," J. Acoust. Soc. Am. **131**(1), 478–491.
Jongman, A., Wayland, R., and Wong, S. (**2000**). "Acoustic characteristics of English fricatives," J. Acoust. Soc. Am. **108**(3), 1252–1263.
Jørgensen, S., and Dau, T. (**2011**). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," J. Acoust. Soc. Am. **130**(3), 1475–1487.
Li, F., and Allen, J. B. (**2009**). "Multiband product rule and consonant identification," J. Acoust. Soc. Am. **126**(1), 347–353.
Li, F., and Allen, J. B. (**2011**). "Manipulation of consonants in natural speech," IEEE Trans. Audio, Speech, Lang. Process. **19**(3), 496–504.
Li, F., Menon, A., and Allen, J. B. (**2010**). "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," J. Acoust. Soc. Am. **127**(4), 2599–2610.
Liberman, A. M., Delattre, P. C., and Cooper, F. S. (**1958**). "Some cues for the distinction between voiced and voiceless stops in initial position," Lang Speech **1**(3), 153–167.
Lobdell, B. E. (**2009**). "Models of human phone transcription in noise based on intelligibility predictors," Ph.D. Thesis, University of Illinois at Urbana–Champaign, Urbana, IL.
Lobdell, B. E., Allen, J. B., and Hasegawa-Johnson, M. A. (**2011**). "Intelligibility predictors and neural representation of speech," Speech Commun. **53**(2), 185–194.
Maniwa, K., and Jongman, A. (**2008**). "Perception of clear fricatives by normal-hearing and simulated hearing-impaired listeners," J. Acoust. Soc. Am. **123**(2), 1114–1125.
Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (**2008**). "Phoneme representation and classification in primary auditory cortex," J. Acoust. Soc. Am. **123**(2), 899–909.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**(2), 338–352.
Phatak, S., Lovitt, A., and Allen, J. B. (**2008**). "Consonant confusions in white noise," J. Acoust. Soc. Am. **124**(2), 1220–1233.
Régnier, M. S., and Allen, J. B. (**2008**). "A method to identify noise-robust perceptual features: application for consonant /t/," J. Acoust. Soc. Am. **123**(5), 2801–2814.
Remez, R., Rubin, P., Pisoni, D., and Carrell, T. (**1981**). "Speech perception without traditional speech cues," Science **212**(4497), 947–949.
Shadle, C. H., and Mair, S. J. (**1996**). "Quantifying spectral characteristics of fricatives," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 1521–1524.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**(5234), 303–304.
Singh, R., and Allen, J. B. (**2012**). "The influence of stop consonants perceptual features on the Articulation Index model," J. Acoust. Soc. Am. **131**(4), 3051–3068.
Soli, S. (**1981**). "Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation," J. Acoust. Soc. Am. **70**(4), 976–984.
Stevens, K. N., and Blumstein, S. E. (**1978**). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**(5), 1358–1368.
Stevens, K. N., Blumstein, S. E., Glicksman, L., Burton, M., and Kurowski, K. (**1992**). "Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters," J. Acoust. Soc. Am. **91**(5), 2979–3000.
Wang, M. D., and Bilger, R. C. (**1973**). "Consonant confusions in noise: A study of perceptual features," J. Acoust. Soc. Am. **54**(5), 1248–1266.
Whalen, D. H. (**1981**). "Effects of vocalic formant transitions and vowel quality on the English [s]-[ŝ] boundary," J. Acoust. Soc. Am. **69**, 275–282.
Whalen, D. H. (**1991**). "Perception of the English /s/ /ʃ/ distinction relies on fricative noises and transitions, not on brief spectral slices," J. Acoust. Soc. Am. **90**(4), 1776–1785.