Perceptual weighting of acoustic cues for accommodating gender-related talker differences heard by listeners with normal hearing and with cochlear implants

Matthew B. Winn, and Ashley N. Moore

Citation: The Journal of the Acoustical Society of America **148**, 496 (2020); doi: 10.1121/10.0001672

View online: https://doi.org/10.1121/10.0001672

View Table of Contents: https://asa.scitation.org/toc/jas/148/2

Published by the Acoustical Society of America

CALL FOR PAPERS
JASA
THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA
Special Issue:
Machine Learning in Acoustics

JASA ARTICLE

Perceptual weighting of acoustic cues for accommodating gender-related talker differences heard by listeners with normal hearing and with cochlear implants

Matthew B. Winn[1,a)] and Ashley N. Moore[2]

[1]*Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, Minnesota 55455, USA*

[2]*Department of Speech & Hearing Sciences, University of Washington, Seattle, Washington 98105, USA*

ABSTRACT:

Listeners must accommodate acoustic differences between vocal tracts and speaking styles of conversation partners—a process called normalization or accommodation. This study explores what acoustic cues are used to make this perceptual adjustment by listeners with normal hearing or with cochlear implants, when the acoustic variability is related to the talker's gender. A continuum between /ʃ/ and /s/ was paired with naturally spoken vocalic contexts that were parametrically manipulated to vary by numerous cues for talker gender including fundamental frequency (F0), vocal tract length (formant spacing), and direct spectral contrast with the fricative. The goal was to examine relative contributions of these cues toward the tendency to have a lower-frequency acoustic boundary for fricatives spoken by men (found in numerous previous studies). Normal hearing listeners relied primarily on formant spacing and much less on F0. The CI listeners were individually variable, with the F0 cue emerging as the strongest cue on average. © *2020 Acoustical Society of America*. https://doi.org/10.1121/10.0001672

(Received 27 February 2020; revised 31 May 2020; accepted 14 July 2020; published online 3 August 2020)

[Editor: Benjamin V. Tucker] Pages: 496–510

I. INTRODUCTION

Considering the number of different talkers and acoustic environments encountered on an everyday basis, it is remarkable that speech perception is such an obvious and accessible skill for most listeners with normal hearing (NH). To efficiently perceive speech sounds, listeners must maintain some amount of perceptual equivalence across a wide range of variability in voice acoustics across talkers; even though one talker's /s/ is acoustically distinct from the next talker's /s/, there is a sense that they function equivalently in speech. When a listener successfully recovers the correct phoneme despite inter-talker differences, they have *accommodated* or *normalized* the acoustic-phonetic variability. Talker gender is a major source of this variability, with women producing higher fundamental frequencies and higher-frequency vowel formants (Hillenbrand *et al.*, 1995), as well as higher spectral peak frequencies for fricative consonants (Jongman *et al.*, 2000) compared to the corresponding acoustic properties produced by men. There are numerous other factors that influence phonetic acoustics and corresponding perception, including phonetic context (e.g., rounded or unrounded lip posture), speech rate (Miller, 1981; Jaekel *et al.*, 2017), and stable acoustic properties of the sound environment (Stilp *et al.*, 2016). Contextual effects are ubiquitous in speech perception (Stilp, 2020), yet the perceptual mechanisms underlying the accommodation of gender-related acoustic differences remain unknown. In this study, we focus on the acoustic cues that listeners use to accommodate differences between women and men, because these differences are widely encountered and among the largest acoustically.

In this study we avoid the term *normalization*, as that concept commonly involves a much more complicated set of accommodations across a whole inventory of sounds (i.e., an entire set of vowels in a dialect), whereas the current study focuses primarily on accommodation of a narrow set of phonemes where the acoustic signatures of gender are well understood. This study further explores whether the acoustic cues used by listeners with NH differ from those used by listeners with cochlear implants (CIs). The latter population is of particular interest because individuals with CIs have notoriously poor frequency perception, which should seem to be an obstacle in the task of accommodating differences in talker gender, since frequency-based cues (formant frequencies, fundamental frequency) are among the more intuitive cues that should be useful, based on the research reviewed below.

The current study was designed to address the question of which gender-related acoustic cues have the biggest effect on phonetic categorization shifts observed in previous literature that used wholesale changes in talkers [specifically, the study done by Winn *et al.* (2013), to be discussed later]. The size of cue-driven perceptual shifts is being used to infer which cues have the largest effect on changes in classifying fricatives in particular, since they have been commonly studied in this branch of the literature.

[a)]Electronic mail: mwinn@umn.edu

A. Acoustic cues relating to talker gender

Fundamental frequency (F0, colloquially "pitch") is likely the most familiar and obvious cue to talker gender; among adults, there is roughly an octave difference between the voices of women and men, although this is variable across language communities and cultures (Liberman, 2013; Andreeva *et al.*, 2014). Because voice pitch can be so variable, it makes sense that there are other acoustic properties that are relevant when identifying a talker's gender. Among the more reliable differences between women and men is vocal tract length (VTL), which emerges acoustically as differences in the formant (resonant) frequencies of the voice; lower formant frequencies are perceived as indicating large body dimensions usually associated with masculinity (van Dommelen and Moxness, 1995). Although the typical gender-related difference in F0 (roughly 100%) is proportionally larger than the corresponding typical differences in formant spacing [roughly 15%–20%; Fant (1966) and Hillenbrand *et al.* (1995)], the VTL cue is preferred over F0 by NH listeners when differentiating between talkers (Barreda and Nearey, 2013), consistent with the notion that talkers have relatively more flexibility to change their F0 than to change the dimensions of their vocal tract. Hillenbrand and Clark (2009) and Fuller *et al.* (2014) both suggest that reliable perception of a change in talker gender is driven by a concurrent change in F0 and VTL, with either cue usually being not entirely effective on its own. Other cues for talker gender include voice breathiness, carried by a complex group of acoustic cues (Maryn *et al.*, 2009) that could be characterized by relative amplitudes of the first few harmonics, or the balance of spectral energy in low- and high-frequency regions (Skuk and Schweinberger, 2014).

El Boghdady *et al.* (2019) showed that better auditory perception of a talker's acoustic properties (F0 and VTL) is associated with better perception of speech perception in competing noise. Examining the perception of voice characteristics is therefore not only worthwhile for its role in understanding how listeners accommodate intra-talker variability, but also in how listeners succeed in everyday environments.

B. Phonetic accommodation of talker gender

Parallel to the direct identification of talker gender is the aforementioned need to *accommodate* the resulting acoustic differences that affect the acoustics of all the phonemes produced by women and men. Phonetic normalization occurs when phonemes are categorized with regard to the acoustic space of the specific talker rather than by absolute acoustic dimensions; without this process, confusions are likely to occur. An analogy can be made with F0; an F0 of 160 Hz might be considered high for a man's voice but low for a woman's voice, so the judgment of whether it is "high" or "low" must be made in context of the particular voice. Normalization is frequently studied in the context of vowel space or dialects, but for the purpose of normalizing to talker gender specifically, a commonly used example is the difference between /ʃ/ (as in "shock") and /s/ (as in "sock"), since it is a well documented channel for expressing one's gender [cf. Munson *et al.* (2006) and Munson (2011)].

Although the phonemes /ʃ/ and /s/ are acoustically complex (McMurray and Jongman, 2011), they could be described in layman's terms simply as "lower-" and "higher-" frequency noises, respectively. The strongest frequencies in /s/ are higher than those in /ʃ/, so given a range of frequencies spanning from /ʃ/ to /s/, there is some boundary below which we would perceive /ʃ/, and above which we would perceive /s/. However, for a woman's voice, we would expect that all the frequencies would be shifted upward relative to a man's voice. Accordingly, the *perceptual* boundary between the two phonemes for a woman's voice should also be shifted up, or else one might confuse a woman's /ʃ/ for a man's /s/. Listeners with normal hearing indeed adjust their categorization boundary consistent with these expectations (Mann and Repp, 1980; Johnson *et al.*, 1999; Winn *et al.*, 2013). The /ʃ/-/s/ contrast is explored here not because it is a particularly important one for lexical distinctions but rather because it is a channel through which gender differences are expressed in voice acoustics. When a listener shows differences in the labeling of the sounds intermediate to a male /ʃ/ and a female /s/, those differences have been interpreted as demonstrating accommodation to talker gender. As the data from the current study will show, these perceptual patterns might also arise because of perception of apparent talker size independently of gender.

Although the difference between /ʃ/ and /s/ is most commonly described as differences in spectral peak frequencies, there are other frameworks as well. Spectral contrast has been proposed as a ubiquitous mechanism of auditory perception that influences phonetic categorization (Stilp *et al.*, 2015; Stilp, 2020) and which also has been invoked as a mechanism to distinguish /ʃ/ and /s/ (Hedrick and Carney, 1997). This property can be expressed as relative amplitude of adjacent consonant and vowel segments within the F3 / F4 region. Other studies express the acoustic difference in terms of spectral center of gravity [cf. Chodroff and Wilson (2020)], which is the magnitude-weighted average of the frequencies that are present in the spectrum, which has the advantage of being computationally simple and a reliably good separator of the phonemes, but the disadvantage of having no clear correlate in the auditory system.

C. The perception of sound through cochlear implants

The issue of talker gender perception and phonetic perception in general is especially interesting in the context of individuals who use cochlear implants (CIs), which are neural prostheses that directly stimulate the auditory nerve to restore a sensation of hearing to people for whom traditional hearing aids are insufficient. They can provide transformational benefit for listeners who might otherwise be devoid of hearing, and are therefore considered to be a tremendously successful medical achievement. However, there are some severe limitations that still exist. In particular, the signal provided by the CI results in poorer clarity of frequency cues. Both of the main mechanisms of frequency coding in the typical-hearing ear (the rate of stimulation and the place of stimulation) are severely compromised in a CI. With some exceptions, CIs generally supply constant-rate pulsatile stimulation that does not change to reflect the rate of input frequency (periodicity), except for crude amplitude modulation coding [for a discussion of how this affects perception of voice pitch, see Gaudrain and Başkent (2018)]. Place of cochlear stimulation is therefore the primary mechanism of frequency coding in CIs, but it is severely limited because of the relatively small number of distinct places of tonotopic stimulation sites (12 to 22), compared to the thousands of tonotopically arranged inner hair cells. Furthermore, even at those distinct sites of electrode placement, the specificity of neural activation from a CI is very poor, because electrical stimulation spreads to activate a much wider region of the cochlea than what would be elicited in acoustic stimulation (Chatterjee and Shannon, 1998; Boëx *et al.*, 2003). These limitations, along with the possibility of neural atrophy, incomplete cochlear insertion [resulting in upward shifting of all frequencies; Holden *et al.* (2013) and Landsberger *et al.* (2015)], variable distances between electrode and neural populations (DeVries *et al.*, 2016), and severely compressed dynamic range (Zeng *et al.*, 2002) present stark challenges to the CI listener who is attempting to recover fine-grained acoustic-phonetic detail in the speech signal.

In light of the limitations in signal quality described above, it is unsurprising that CI listeners tend to have difficulty in perceiving voice pitch (Gaudrain and Başkent, 2018) and have difficulty in tasks where voice pitch cues should be relevant (Shannon *et al.*, 2004; Luo and Fu, 2006). Additionally, CI users have difficulty perceiving other spectral cues in speech, such as those that differentiate consonant place of articulation (Munson *et al.*, 2003; Winn and Litovsky, 2015) and VTL (Fuller *et al.*, 2014; Gaudrain and Başkent, 2018).

Winn *et al.* (2013) began with the premise that CI users have poor frequency resolution and hypothesized that they should therefore show a diminished ability to accommodate talker gender when categorizing phonemes, since gender-related acoustic differences are cued spectrally. Their experiment featured a continuum of fricative sounds spanning the acoustic space between /ʃ/ and /s/ embedded in words with /i/ and /u/ vowels, producing the words “see, sue, she, shoe”; the vowel contained the acoustic cue to talker gender, as it was uttered by one of two women or two men. Each fricative-vowel combination was heard multiple times and the categorization of the reported onset consonant was analyzed in order to identify the perceptual boundary between the two phonemes. Surprisingly, CI listeners demonstrated phonetic accommodation of talker gender by exhibiting phonetic boundaries at lower-frequency sounds for male voices, with an effect size that was comparable to that shown by NH listeners. The study also analyzed accommodation of vowel-related cues such as lip rounding as well as visual cues to gender and coarticulatory compatibility of the consonant and vowel combination; these perceptions are not under investigation in the current paper. Considering the limitations of the CI in transmitting acoustic details, it remains unclear what cues CI listeners could be using to accomplish the task of accommodating gender-related acoustic cues in vowels that drive shifts in perception of adjacent fricative. The current study was designed to address that open question by manipulating specific acoustic cues that are likely to play a role, based on the literature reviewed below.

D. Perception of talker gender by listeners with cochlear implants

Despite challenges in perception of the cues thought to be responsible for gender perception, CI listeners are able to both identify talker gender and also accommodate gender-related acoustic-phonetic differences in fricatives. Fu *et al.* (2005) found that F0 differences between women and men are crucial to this ability, with small differences (on the order of 10 Hz) resulting in poor performance. Kovačić and Balaban (2009) found that talker gender identification in CI listeners is highly variable, with only half of their 20-participant sample able to reliable perform the task.

Detailed analyses of the perceptual strategy for gender perception in CI listeners reveals potential sources of difficulties and potential for the presence of atypical listening strategies or cue weighting. Data from Fuller *et al.* (2014) suggest that CI listeners rely more heavily on F0 cues compared to orthogonal VTL cues for direct gender identification. In the same experiment, NH listeners relied on a combination of F0 and VTL cues to identify talker gender. These results were somewhat surprising given the relative difficulty of CI listeners in perceiving pitch. However, VTL is also understandably challenging to ascertain, as formant structure is generally more difficult to perceive by CI listeners (Winn *et al.*, 2012; Winn and Litovsky, 2015; Gaudrain and Başkent, 2015). Follow-up work by Gaudrain and Başkent (2018) using stimuli similar to those used by Fuller *et al.* (2014) support the notion that the F0 changes within the range representative of gender differences are more discriminable than corresponding changes in VTL.

E. Research questions and hypotheses

The current study was designed to address two questions: (1) What acoustic cues contribute most to phonetic accommodation of talker gender when NH listeners categorize fricatives? (2) Do CI listeners use the same cues as those used by listeners with NH? To address these questions, a cue-weighting study was conducted in which four acoustic cues for talker gender were parametrically varied in vowels appended to each member of a continuum of fricative consonants; the gender-related acoustic cues were manipulated in the vowels, and the perceptual influence of those cues was operationally defined as the change in categorization of the fricatives. Specifically, the increased tendency to label the fricatives as /s/ instead of /ʃ/ was interpreted as a signature of having a lower-frequency perceptual boundary, which is traditionally seen as reflecting perception of a masculine (i.e., larger) talker. The stimuli were designed not to test direct gender identification (e.g., "Is this a woman or a man?") but instead were designed to measure the phonetic accommodation that should ensue after a talker ascertains the acoustic space of the talker (e.g., "This is a woman, I therefore expect the /s/ phoneme to have higher frequencies than if it were spoken by a man").

Based on the literature reviewed above, it was hypothesized that (1) there should be a hierarchy of cue importance for NH listeners that prioritizes VTL, considering the reasonable expectation that difference in vowel resonance frequencies would scale commensurate with resonance frequencies for fricatives. (2) The impoverished auditory input associated with a CI might require an alternative strategy of phonetic accommodation [cf. Winn *et al.* (2012) and Winn *et al.* (2013)]. On the basis of prior studies that suggested a substantial role of spectral contrast in phonetic context effects [cf. Stilp (2020)], we hypothesized that the phonetic accommodation effect in NH listeners would be further affected by the relative amplitude of energy in the vowel nearest the spectral peak in the fricative. Although F0 has been identified as an important contributor to direct identification of gender, we did not hypothesize that it would be related to phonetic accommodation in NH listeners since the laryngeal mechanism responsible for voice pitch is physically independent of the VTL/spectral shape. However, F0 has been shown to indirectly affect vowel perception (Barreda and Nearey, 2012) and could be interpreted as a surrogate or proxy cue for large vocal tract size, so it was feasible for this cue to also influence phonetic categorization in the same direction as VTL. Considering the multiplicity of acoustic cues available to NH listeners and the variable degradation of some of those cues in CI listeners, it is understandable that acoustic-phonetic cue weighting for phonetic contrasts tends to be atypical in CI listeners [for vowels and consonant voicing: Winn *et al.* (2012); for consonant manner of articulation: Moberly *et al.* (2015); for consonant place of articulation: Winn and Litovsky (2015)]. We therefore hypothesized that acoustic cue weighting would be markedly different between these listener groups for gender-related phonetic accommodation.

II. METHODS

A. Participants

Participants included 21 adult listeners (29.5 years mean age; 10 female; 11 male) with normal hearing who demonstrated pure-tone thresholds ≤20 dB hearing level at octave frequencies from 250 to 8000 Hz bilaterally. There were also 19 adult cochlear implant users (22–87 years of age; 61 years mean age, 13 women, 6 men), of whom all but 1 were post-lingually deafened. Two additional pre-lingually deaf participants with CIs were tested but were unable to complete the task because they reported that none of the sounds were discriminable; their contributions were excluded from the data set reported in the results. Table I contains demographic information for individual CI users. Although CI participants in this sample were generally older than the listeners in the NH group, the ability under investigation—gender-related phonetic accommodation—was previously found to be virtually completely intact even in older CI users (Winn *et al.*, 2013). All listeners were native speakers of American English and gave informed consent to participate in this study. Procedures were approved by the University of Washington Human Subjects Division and the Institutional Review Board at the University of Minnesota.

TABLE I. Demographics of CI participants.

Listener	Sex	Age	Device type	Implanted ear(s)	Etiology of deafness	CI Exp. (years since 1st CI)
C101	F	54	MedEl Sonnet	Bilateral	Sudden SNHL	5
C102	F	64	Cochlear N6	Right	Idiopathic	2
C104	M	64	AB Naida Q70	Bilateral	Ototoxicity	15
C105	F	47	Cochlear N6	Bilateral	Progressive SNHL	8
C106	M	87	AB Naida Q90	Bilateral	Noise-related SNHL	30
C108	F	73	Cochlear N7	Right	Progressive SNHL	8
C109	M	47	AB Naida Q70	Left	Auditory neuropathy	2
C110	M	78	Cochlear N6	Bilateral	Progressive SNHL	14
C112	F	79	MedEl Sonnet	Left	Unknown	2
C113	M	72	AB Naida Q90	Bilateral	Progressive SNHL	10
C116	F	61	AB Naida	Right	Rheumatic fever	22
C117	M	66	AB Naida Q70	Bilateral	Auditory neuropathy	7
C118	F	30	Cochlear N7	Bilateral	Sudden SNHL	7.5
C119	F	22	Cochlear N7	Bilateral	Unknown	17
C123	F	60	AB Harmony	Left	Genetic	9
C137	F	59	Cochlear Kanso	Bilateral	Unknown	2
C138	F	60	AB Naida	Bilateral	Unknown	27
C139	F	61	AB Naida Q70	Bilateral	Genetic	7
C141	F	73	AB Naida Q70	Right	Genetic	6.5

B. Stimulus overview

The stimuli in this experiment were a modified subset of those used in a prior study by Winn *et al.* (2013). In brief, there was a fully crossed combination of two vowel environments (/i/ and /u/), two levels of VTL (formant spacing that was typical of either a woman or a man), two levels of F0 (typical of either a woman or a man), and three levels of spectral contrast, each for two source voices [each acoustic dimension was modified using natural recordings of a woman and a man; both talkers were used in the original study by Winn *et al.* (2013)]. The acoustic properties of each manipulated cue were set to match the difference that naturally emerged in the original recordings, with the exception of spectral tilt, which was set according to long-term average spectral differences of a different set of recordings of women and men so that these effects were not limited only to the voices used here. The 2 × 2 × 2 × 3 × 2 (vowel:VTL:F0:tilt:source-voice) combination resulted in a total of 48 distinct vocalic contexts. Each of these contexts were appended to an eight-step fricative continuum that ranged from a low-frequency /ʃ/ to a high-frequency /s/, described below. In total there were 384 distinct stimuli; each unique stimulus was presented four times to each listener.

C. Fricative continuum synthesis

The fricative continuum in this study was modeled off of the first eight steps of the nine-step continuum manually created by Winn *et al.* (2013). The ninth step was omitted since steps 7 and 8 were both perceived as /s/ nearly 100% of the time, rendering step 9 redundant. There were other alterations to improve the replicability and standardization of the fricatives. Instead of the manual method used in the 2013 study, the sounds were synthesized using a scripted procedure in PRAAT [version 6.026; Boersma and Weenink (2017)] by filtering and adding narrowband noises filtered from white noise. Each continuum step contained three spectral peaks that differed by central frequency, bandwidth, and relative amplitude [for more details, see the methods used by Winn (2020) and the PRAAT script in the supplemental materials of this paper[1]]. The continuum gradually transitioned between /ʃ/ and /s/ and was regarded by the experimenters and listeners as highly natural when appended to naturally spoken vowels. Each fricative was 200 ms long, with a 140 ms onset ramp and 40 ms offset ramp in its amplitude envelope. All fricatives were equalized for root-mean-square intensity. Parameters of the fricative continuum as well as spectra of each continuum step are illustrated in Fig. 1 and available in a table in the supplemental materials.[1]

The primary acoustic distinctions across the /ʃ/-/s/ continuum were the center frequencies and relative bandwidths of the noise peaks, with noise peak bandwidth also altered, although in a less noticeable fashion. At the low-frequency /ʃ/ end, the center frequencies of the spectral peaks were 2930, 6130, and 8100 Hz; the lowest peak was 4 dB higher than the middle peak and the highest peak was 2 dB lower than the middle peak, creating a downward sloping spectral tilt. At the high-frequency /s/ end of the continuum, the peaks were centered at 5725, 7900, and 9485 Hz; the lowest peak was 6 dB lower than the middle peak and the highest peak was 6 dB higher than the middle peak, creating a steep upward spectral tilt. The bandwidth of each peak was attenuated at 48 dB/octave at the /ʃ/ end and this narrowed to 96 dB/octave at the /s/-end of the continuum. All frequency filtering was done using custom filters that operated on the FFT spectrum object and then were inverted to create sounds with appropriately modified spectra.

Following spectral filtering, the amplitude envelope was modified to create a 150 ms onset ramp and 40 ms offset ramp over the entire 200 ms duration of the fricative. This envelope was applied uniformly across the entire continuum. Low-frequency envelope distortions resulting from the intensity modification were spectrally remote from the fricative energy itself and were not perceptible.

The /s/-end of the fricative continuum contained substantial energy in frequency regions that are not sampled by cochlear implant speech processors (i.e., above 8000 Hz). This limitation is not expected to affect the results of the current study because the main effect of interest was how gender-related cues in the vocalic context affected

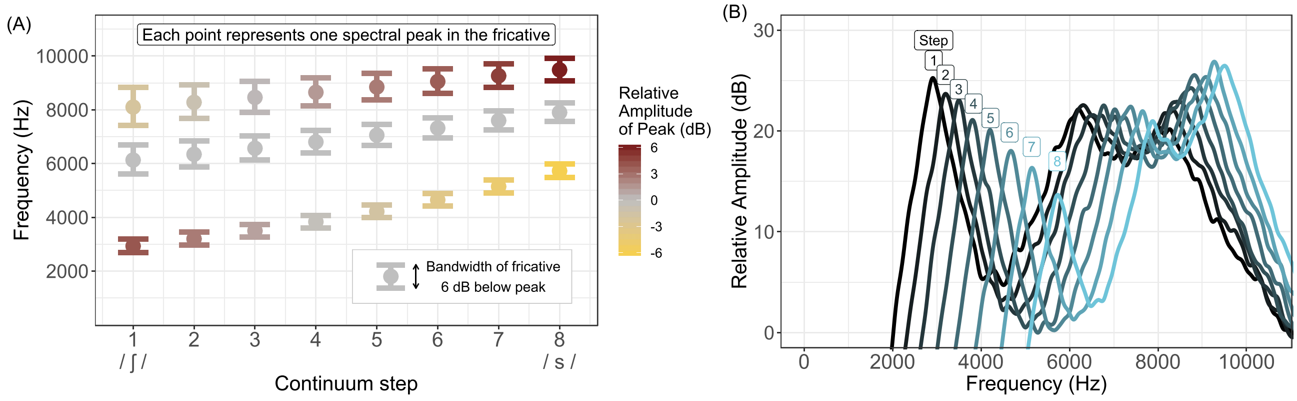

FIG. 1. (Color online) Parameters of the fricative continuum. (A) Frequencies, bandwidths, and relative amplitudes of three spectral peaks for each of eight continuum steps. (B) Corresponding spectra of the fricative sounds.

perception of fricative continuum, rather than an examination of fricative perception itself. That is, should the lack of full /s/ bandwidth be perceptible, it should be equally perceptible in the context of a male voice as well as a female voice appended to the fricative.

D. Vocalic contexts

Each step of the fricative continuum was pre-appended to a digitally altered version of the vowel /i/ or /u/; the vowel contains the acoustic properties that would convey the gender of the talker, since the synthetic fricatives were not informative for talker identity in themselves. These vowels were recordings of naturally spoken utterances from two native English talkers (one male and female) saying the words "see" and "sue." The corresponding vowels from /ʃ/-onset words were excluded to limit the number of stimuli, and because the associated effect of formant transition was the weakest of all previously tested by Winn *et al.* (2013). Furthermore, pilot listening suggested that the vowels extracted from /s/-onset words sounded natural with the entire range of fricatives, whereas the vowels extracted from /ʃ/-onset words sounded natural only at the /ʃ/ end of the continuum. Two vowels /i/ and /u/ were used primarily for the purpose of reducing monotony of stimuli (i.e., to play four words instead of two), and secondarily because there is an additional contextual effect of vowel environment that could potentially be worthy of investigation (Mann and Repp, 1980; Winn *et al.*, 2013). Additionally, the use of multiple vowels is also useful in avoiding the use of spectral information for talker identity that happens to be specific to a single vowel, as noted by Barreda (2016), who demonstrated bias toward perceiving taller talkers when presented with high-back vowels.

1. Source voices

Perceived talker size (and therefore potentially talker sex or gender) is affected by perception of inherent spectral and source characteristics of vowels (Barreda, 2017). It was therefore desirable to see if any residual voice-source cues could exert any influence over perceptual judgments separately of the main cues such as VTL and F0. Each of the aforementioned acoustic cues for gender (F0, VTL, spectral tilt) were independently manipulated in vowels spoken by a woman and by a man.

2. Fundamental frequency (F0)

F0 was manipulated with PRAAT using the PSOLA (pitch-synchronous overlap-add) method. The F0 contour of the woman's voice was imposed onto the vowel spoken by the male talker and vice versa. The stimuli without any F0 shift were processed through the same algorithm (original pitch contour replaced by a replica of itself) to ensure that any artifacts of processing were applied equally to all stimuli. The average F0 for the "male" contour was 104 Hz, and for the "female" contour was 208 Hz. The contour maintained natural dynamics and had a shallow "U" shape, rather than being flat.

3. VTL

VTL is conveyed by the spacing of the formant frequencies in the vowel. In order to transform the formant spacing, we first obtained the vowel-specific ratio of female to male frequency values for F1, F2, and F3 using the original recordings and also by using the Hillenbrand *et al.* (1995) vowel database as a guide. For shifting the male voice to a female voice, the VTL of /i/ vowel was shifted by a factor of 1.17, which was close to the ratio of 1.19 for F2 in this vowel in the Hillenbrand *et al.* database. In reverse, the female VTL for /i/ was shifted by a factor of 0.84 to become more masculine. The VTL shift for the /u/ vowel was more substantial, with the male-female multiplication factor of 1.28 (slightly smaller than the database value of 1.44), and female-to-male multiplication factor of 0.78. Reciprocal multiplication factors were applied to both the female and male voices in the study, rather than having a single natural endpoint. Therefore, the "female" or "male" vocal tract configurations were orthogonally crossed with "female" and "male" source voices. These values for VTL modification were not used because they represent a population-level difference across the sexes but rather because they are the actual proportional differences between the talkers recorded and used for stimuli for previous studies.

To shift the formant spacing by a factor of X, we overrode the sampling frequency by a factor of X, and then increased the sound's duration (using the PSOLA method) by a factor of $1/X$. For example, to make a vowel sound like it was spoken by a larger (more "male") talker, a sampling rate of 44 100 Hz was changed by a factor of 1/1.2 to be 36 750 Hz (which made the talker sound larger and also lengthened the audio by a factor of 1.2), and then the original duration was restored by applying the PSOLA algorithm using a duration factor of (1/1.2, or 0.833). Only the vowel (not the fricative) was processed by this resampling and time-warping method. Figure 2 illustrates an example of the shifting of formant frequencies for the same vowel /i/ spoken by the female talker in this study. In that figure, the formant tracks are overlaid on the spectrograms of vowels that have undergone the VTL shifting.

After modifying the formant frequencies using the procedure explained above, it was observed that there were residual differences in the high frequency regions of the stimuli that were mathematically inevitable, since the entire spectrum was shifted (not just F1, F2, and F3). These differences were mainly in the amount of spectral energy above 3200 Hz, which are notable because they might manifest perceptually as differences in local spectral contrast in the region of the fricative. Because high-frequency spectral contrast is known to influence consonant categorization independent of any VTL manipulations (Lotto and Kluender, 1998), these differences needed to be eliminated. A linearly sloped filter in the frequency domain

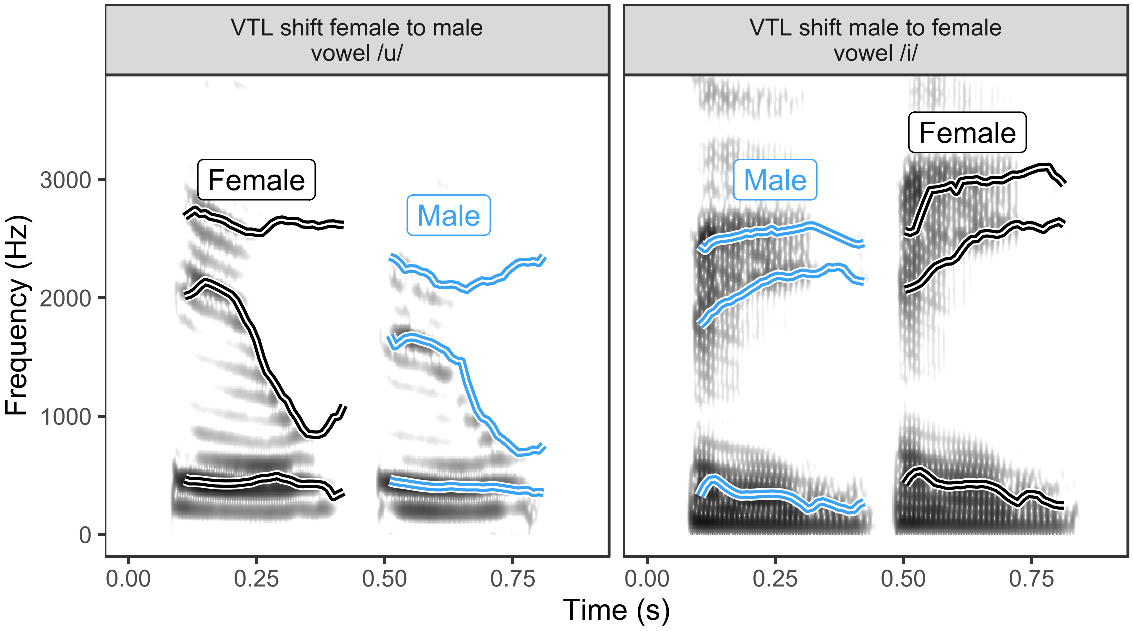

FIG. 2. (Color online) Shifting of formant tracks for vowels that have undergone a VTL shift from female to male (left panel) for the /u/ vowel, and a VTL shift from male to female for the /i/ vowel (right panel).

was used, which was adjusted for each VTL-manipulated sound until both the female and male stimuli had virtually equivalent spectral tilt (defined generally as the slope in intensity from F1 to F3). Then, all frequency energy above 3200 Hz was equalized in both stimuli, by first low-pass filtering the modified sounds, and adding them with a uniform high-passed portion of the original sound. Consequently, both original and manipulated sounds had the original high-frequency portion maintained, and VTL-adjusted formants only differed below 3200 Hz (which included F1, F2, and F3). Figure 3 illustrates the sequential process of formant shifting and subsequent normalization of high-frequency energy while lower-frequency formants were manipulated.

4. Spectral contrast

The spectral contrast feature in the current stimulus set corresponded to a gain or attenuation filter targeting energy in the frequency region corresponding to the frequency peak step 4 of the fricative continuum, which had previously been found (Winn, 2020) to be rather ambiguous. Using a custom filter function in PRAAT (implemented linearly in the FFT domain), the filter applied 12 dB gain or attenuation at the center frequency, with the spectrum changes linearly tapering to zero at 1200 and 8000 Hz. The most substantial changes in frequency filtering were within 2 and 6 kHz; that range will be used in this paper as a shorthand for this particular filtering procedure. In addition to the stimuli that were given a positive or negative filter gain, the remaining 1/3 of stimuli were left unaltered. Figure 4 illustrates the different amounts of spectral filtering layered against the spectrum of the fricative in the center of the continuum.

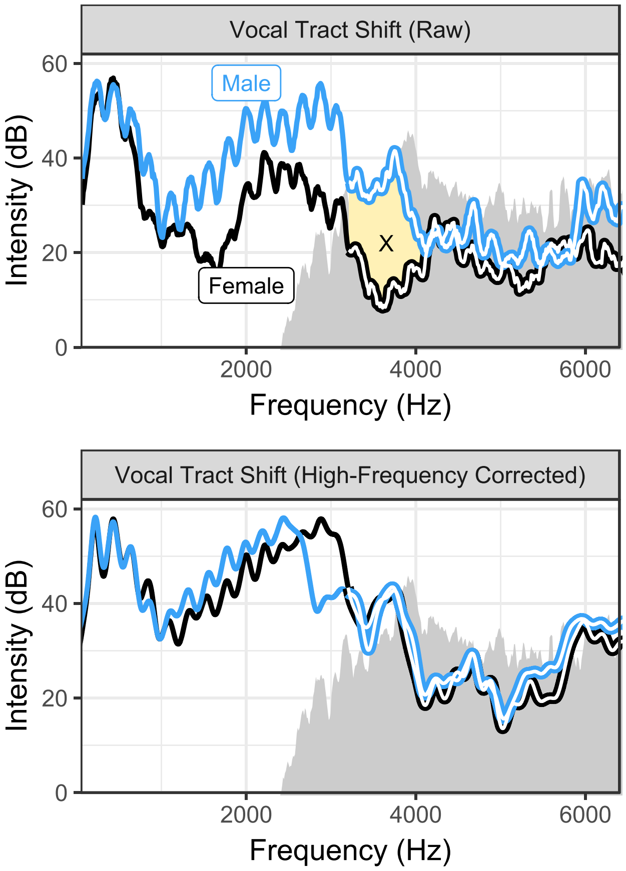

FIG. 3. (Color online) Top panel illustrates unintended area of spectral contrast between the vowel (barred line) and fricative (filled gray region), where that region of spectral contrast is highlighted with yellow and marked with an X. Bottom panel illustrates two sounds that maintain the VTL shift but with correction of this unintended spectral contrast by equalizing the spectral energy level in the vowel in the frequency region above 3200 Hz.

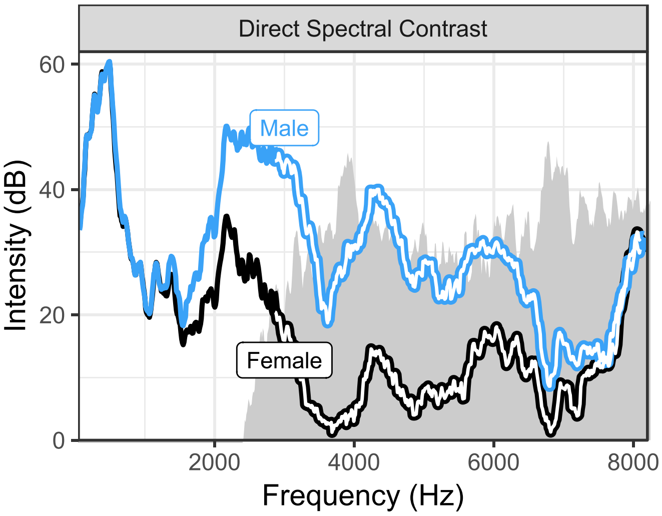

FIG. 4. (Color online) Illustration of the same vowel (lines) with spectral filtering to raise (blue/"male") or lower (black/"female") spectral energy centered at 4000 Hz. A fricative whose lowest spectral peak frequency is 4000 Hz is plotted as the filled gray area for comparison.

E. Procedure

Both NH and CI listeners were presented with a four-alternative forced-choice task. The four selections ("see," "sue," "she," and "shoe") were displayed on a monitor in front of the participant. Stimuli were presented at a comfortable listening level of 65 dBA in the free field from a single loudspeaker that was positioned in front of the participant. After the auditory stimulus was presented, the participant would select what they heard using a computer mouse. Each testing block contained a random half (192) of the full stimulus set, and the presentation of tokens was randomized within each block of the experiment. There were eight total testing blocks which each took approximately 8–12 min each. All testing was conducted in a double-walled sound-treated booth (Interacoustics RE-243). Listeners were not informed until after testing as to the aims of the study. All CI listeners wore their everyday speech processors in the setting that they would normally use for conversation in a quiet room. Their devices were not verified with fitting software by the experimenter, but no participants reported any abnormal device function. Bilateral CI listeners used both processors during testing.

F. Analysis

Binomial logistic regression was used to categorize responses from each participant using open source R software interface (version 3.22, R Development Core Team, 2016), using the lme4 package [version 1.1–12; Bates *et al.* (2015)]. The binomial analysis reflected the idea that responses either began with the fricative /ʃ/ (coded as 0) or /s/ (coded as 1). Steps in the fricative continuum were coded in the model input using indices centered at the hypothetical step number 4.5 (i.e., step 7 was coded as +2.5, step 1 was coded as −3.5). This was chosen so that the default model parameters described the center of the continuum, where the contextual properties exerted strong effects in the raw data, rather than endpoints, where the context effects would be imputed mathematically by warping the proportional space via the logit transform. Contextual factors were contrast coded so that /s/-biasing ("male") cue levels were coded as +0.5 and /ʃ/-biasing ("female") cue levels were coded as −0.5. This resulted in a default model that did not use either gender as a default but rather modeled each fixed effect as a full change from one gender to the other, averaging across the mean of all other fixed effects.

A mixed-effects generalized linear model contained main fixed effects of fricative step, VTL, F0, spectral tilt, original talker, and hearing status. Interactions between main effects were retained in the model if they produced significant improvements in model fit according to the Akaike Information Criterion (Akaike, 1974). There were random effects of intercept, slope (fricative step), VTL, F0, spectral contrast and original voice, each expressed over individual listeners and also over the two listener groups. The prevailing model had the following form:

$$\begin{aligned}&\mathrm{glmer(resp.s \sim step.c + VTL + F0 + sc + orig + Hearing}\\&\quad \mathrm{+step.c : Hearing + VTL : Hearing}\\&\quad \mathrm{+ F0 : Hearing + sc : Hearing + orig : Hearing}\\&\quad \mathrm{+step.c : VTL}\\&\quad \mathrm{+step.c : VTL : Hearing}\\&\quad \mathrm{+(1 + step.c + VTL + F0 + sc + orig | Listener)}\\&\quad \mathrm{+(1 + step.c + VTL + F0 + sc + orig | Hearing),}\\&\quad \mathrm{family = ``binomial'').}\end{aligned}$$

In this model, "resp_s" refers to the outcome measure of responding with a word that begins with /s/ (see or sue), "step.c" is the centered continuum step number, "sc" refers to spectral contrast, and "orig" refers to the gender of the talker in the original recording, i.e., the source voice. Two-way interactions are indicated by a colon; "step.c:VTL" refers to the change in the effect of step.c (i.e., the change in slope) as the VTL is changed.

III. RESULTS

Listeners reliably categorized the fricative continuum into /ʃ/ and /s/ at endpoints, and showed stereotypical psychometric function with identifiable crossover points. Figure 5 shows labeling functions broken down by specific acoustic cues, ordered by strength of effect in the NH listener group. Responses to the "natural voice" are for stimuli where the natural vocal cues were all coordinated to indicate a female or a male voice; this reflects the comparison that should elicit the theoretical maximum context effect. Other panels in the graph reflect labeling functions when changing one cue (e.g., changing from a feminine VTL to a masculine VTL) while averaging over all levels of all other cues. Response functions changed in expected directions when changing F0 and VTL; more /s/ responses were elicited when these cues were more representative of male voices.

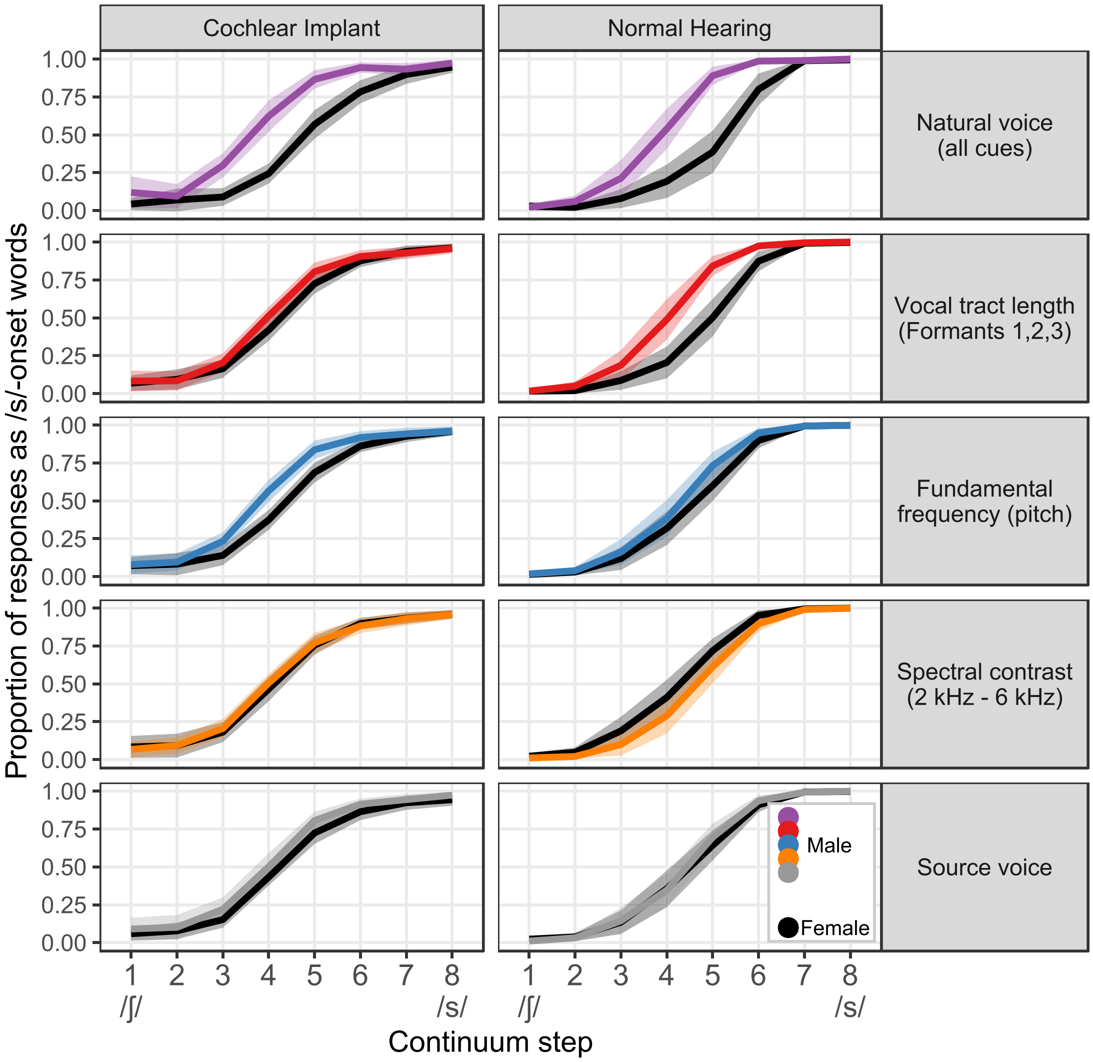

FIG. 5. (Color online) Psychometric functions organized by specific cue to talker gender for listeners with cochlear implants (left column) and with normal hearing (right column). The top row shows responses when vowels contained all consistent cues for talker gender (i.e., no conflicting cues). Each subsequent row shows labeling functions for contrasting levels of a specific cue (e.g., VTL “female” in black versus VTL “male” another color), averaged across all levels of the other cues, including conflicting levels.

The effects of other factors were more variable and less pronounced, to be explained further below.

Table II shows the results of the generalized linear mixed-effects model that was used to account for the log-odds of perceiving /s/, given various combinations of acoustic cues. In the center of the continuum, there was a detectable bias toward hearing /s/ for NH listeners, averaging across all levels of the cues ($p = 0.021$; $\beta 1$ in Table II). There was a strong effect of continuum step for NH listeners, with log odds of /s/ changing by nearly 2 for each single step change across the continuum ($\beta 2$, $p < 0.001$). There were statistically detectable effects of VTL ($\beta 3$, $p < 0.001$) F0 ($\beta 4$, $p < 0.001$). The effect of spectral contrast was also detectable and went in reverse direction to what was expected ($\beta 5$, $p < 0.001$). There was no statistically detectable effect of source voice for NH listeners ($\beta 6$, $p = 0.552$).

For CI listeners, there was no statistical difference in the overall bias at continuum default levels for CI listeners compared to NH listeners ($\beta 7$, $p = 0.905$). The effect of continuum step was statistically smaller for CI listeners compared to NH listeners (i.e., there were shallower psychometric function slopes; $\beta 8$, $p < 0.001$). The effect of VTL was statistically weaker for CI users compared to NH listeners ($\beta 9$, $p < 0.001$). The effect of F0 for CI listeners appeared larger for CI users, but did not reach conventional criterion for statistical difference when comparison against NH listeners ($\beta 10$, $p = 0.191$). The effect of spectral contrast was larger for CI users compared to NH listeners, and in fact changed sign from negative to positive ($\beta 11$, $p < 0.001$). However, the effect of spectral contrast in CI users was not statistically different from zero [$\beta = 0.118$; standard error (st.err) $= 0.113$; $z = 1.043$; $p = 0.297$]. The effect of source voice for CI listeners was statistically different from zero ($\beta = 0.294$; st.err $= 0.129$; $z = 2.28$; $p = 0.022$), but did not reach conventional criterion for statistical difference when comparison against NH listeners ($\beta 12$, $p = 0.205$), for whom the effect was not different than zero.

TABLE II. Results of generalized linear mixed-effects model estimating the factors influencing the perception of /s/ in the context of multiple cues to voice gender, in listeners with normal hearing or with cochlear implants.

	Term	Estimate	std.error	t statistic	p (|z|)
b 1	Intercept	0.240	0.104	2.310	0.021
b 2	Continuum step	1.968	0.128	15.431	< 0.001
b 3	VTL	1.591	0.095	16.661	< 0.001
b 4	F0	0.602	0.091	6.585	< 0.001
b 5	Spectral contrast	−0.773	0.109	−7.094	< 0.001
b 6	Source voice	0.080	0.125	0.640	0.522
Interaction of Hearing (CI) with main effects					
b 7	Intercept: CI	−0.018	0.150	−0.119	0.905
b 8	Continuum step: CI	−0.762	0.184	−4.134	< 0.001
b 9	VTL: CI	−1.291	0.137	−9.447	< 0.001
b 10	F0: CI	0.172	0.132	1.307	0.191
b 11	Spectral Contrast: CI	0.888	0.157	5.658	< 0.001
b 12	Source V: CI	0.229	0.181	1.267	0.205

A. Normalization of direct context effects

A focal point of the current study is the spacing between the psychometric functions corresponding to different levels of each acoustic cue, which represents the influence of gender-related contextual factors on fricative identification. By applying select "male" or "female" acoustic properties orthogonally, the goal in this analysis was to see which cue produces the greatest shift in psychometric functions, which would suggest that it is most responsible for the overall shift in natural voices with the full complement of acoustic cues. We wished to visualize the effects of these cues using the raw rather than modeled data, and to express the effects in proportional space rather than logit space, for the ease of interpreting these effects and contextualizing them across other studies with similarly styled phoneme-categorization functions. A challenge is that psychometric functions for individuals might have different continuum steps where the effect emerges most strongly. Therefore, a normalization method was devised to handle this exact type of variability.

Figure 6 shows an illustration of how context effects were normalized across individuals by centering the continuum indices at the step that elicited the greatest context effect. In this idealized example, the difference in psychometric functions is visualized for two different hypothetical cues labeled X and Y. Cue X (on the top left panel), exhibits a large context effect, observed as more space between the psychometric functions, compared to cue Y in the top right panel. When directly plotting the difference between curves in the panels below the functions, we see a larger peak in correspondence with the difference between the curves on the left; this is the signature of a stronger effect of the contextual vocalic cue. We identified the continuum step that elicited the peak difference between curves and transformed that step index to be 0, flanked with negative and positive steps to the left and right. Using this transformation, a listener whose maximum effect at step 3 has a difference-between-curves function aligned with another listener whose maximum effect is at step 5, rather than averaging one listener's peak with the side band of a different listener's function (which would reduce the estimated effect sizes in a misleading way). In other words, this transformation was done so that large effects across listeners would not be erroneously under-estimated just because they occurred at different continuum steps.

After aligning the differences between curves, the effects on the negative and positive side of the peak were summed (following visual inspection that suggested approximate symmetry on opposite sides of the peaks) so that the continuum is expressed simply as deviation from the center of the listener's personal continuum midpoint. It is important to note that there are mathematical and philosophical differences between this approach and the logit-transform approach taken in the full statistical model described earlier. Whereas the binomial logistic model transforms the probability space to essentially impose a constant intercept effect across the entire continuum (expanding differences at the edges of the continuum), the probability model maintains small context effects at the unambiguous phoneme endpoints (i.e., a clear /ʃ/ or a clear /s/), with the effect of context emerging primarily in the ambiguous center of the continuum. Each approach offers its set of advantages and disadvantages. The binomial model is a mathematically elegant way to handle the data, which are clearly sigmoidal, and which reduce to good approximations of linear lines when transformed to logit space. The probability data on the other hand, are consistent with the experience of relatively weak effect of context for unambiguous phonemes, and strong effects of context when the phoneme is ambiguous.

Using the style of visualization displayed in Fig. 6, average direct effect sizes for all acoustic cue conditions for each group of listeners are shown in Fig. 7, directly illustrating the strength of each cue compared to the others.

Together with Fig. 5, Fig. 7 shows that for NH listeners, the greatest contribution to phonetic accommodation of talker gender appears to be VTL, followed by F0. The effect of original voice was virtually nothing, suggesting that NH listeners did not use any residual acoustic cues other than those that were modified. The effect of spectral contrast went in the reverse direction compared to the hypothesis (this value in Fig. 7 is negative).

For CI listeners, the effect of the full complement of cues in the natural voice was smaller than the same effect in NH listeners, consistent with earlier results (Winn *et al.*, 2013). Across the entire group, F0 was the cue that carried the greatest influence for this phonetic accommodation task. However, the use of specific acoustic cues by CI listeners was generally lower on average compared to the NH group. This was likely due largely to the individual variability across listeners; if only a portion of listeners used a cue very strongly, the lack of use by the rest of the group brought down the average. As will be discussed below, variability among CI listeners was even greater than anticipated; not all CI listeners used F0 for phonetic accommodation, so this cue should not be interpreted as a strong or reliable cue overall.

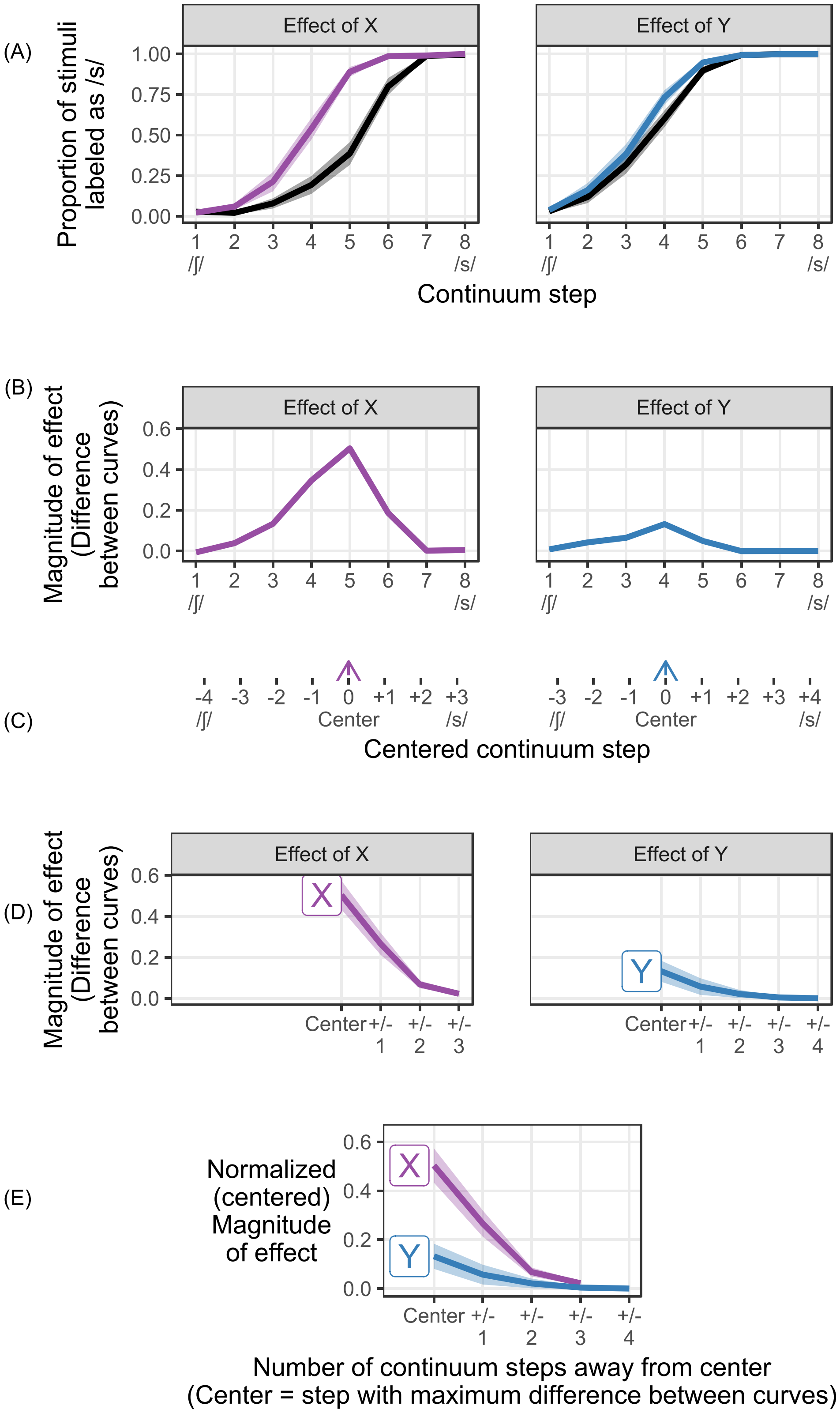

FIG. 6. (Color online) Illustration of how the raw psychometric functions (A) are decomposed into difference-between-curves functions (B), whose continuum steps are then centered on the step that yielded the maximum difference (C). The data are grouped by absolute deviation symmetrical to the central peak and averaged (D). Multiple cues can then be compared on the same panel and arranged according to the magnitude of the direct effect (E).

B. Individual variability

All NH listeners were most influenced by the VTL cue, with 17 of 21 showing F0 to be the second-highest weighted cue. Conversely, there was substantial variability across CI listeners with regard to which cue contributed most to talker accommodation. Individual CI participant cue-strength plots are illustrated in Fig. 8, ordered by magnitude of overall context effect for the natural voices. For participants C110, C105, and C11, the weighting pattern is very similar to what we observed in the NH group, where VTL cues prevailed as the dominant cue. Conversely, C104, C123, C137, and C139

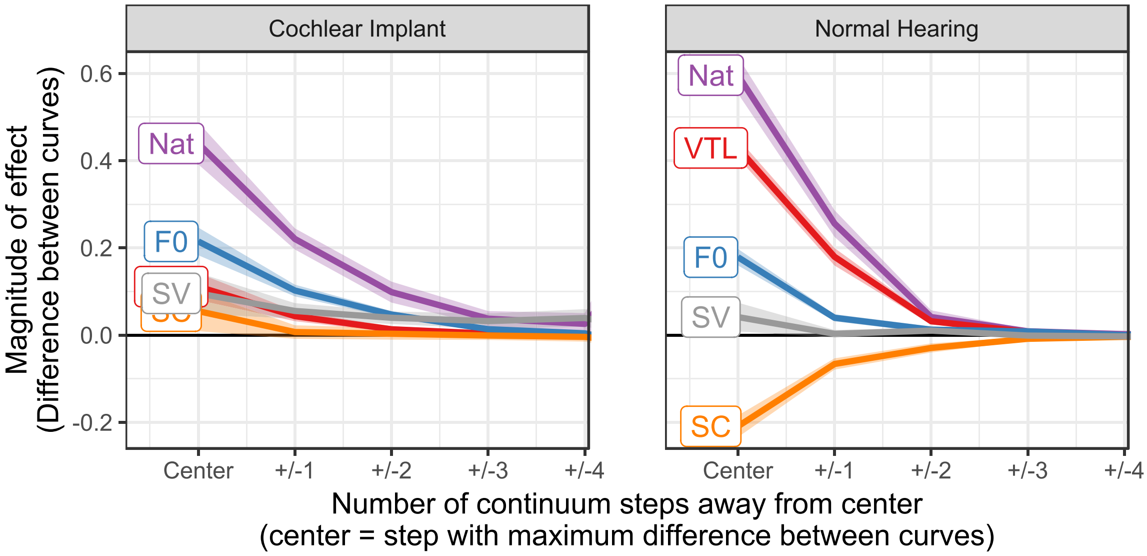

FIG. 7. (Color online) Direct effect of each acoustic cue for CI and NH listeners, plotted as average magnitude of the difference between psychometric functions at continuum steps relative to each listener's centered continuum. Width of ribbons around the data indicate +/−1 standard error of the mean. Nat = natural voice (all cues complementary); F0 = fundamental frequency, SV = source voice; SC = spectral contrast.

demonstrated greatest reliance on F0 cues. Some participants (C117 and C138) demonstrate very little phonetic accommodation even for the fully natural voice. Other listeners (C116, C119) demonstrated an accommodation effect for the natural voice but no clear use of acoustic cues when those cues were in isolation. Surprisingly, there were numerous CI listeners who showed greatest influence of the *source voice*, meaning that none of the actual modified acoustic cues drove their perception as much as the simple factor of which voice was the one being manipulated.

IV. DISCUSSION

There are multiple acoustic differences between the voices of women and men, some of which might interact with acoustic cues used for phoneme recognition. Listeners therefore must identify phonemes within the context of the voice they are hearing, using a variety of acoustic cues to guide expectations for where the acoustic-phonetic boundaries lie between phonemes. In this study, the influence of these acoustic cues was estimated by measuring the impact on fricative perception resulting from orthogonal acoustic manipulation of the gender-related cues in the subsequent vowel.

Both CI listeners and NH listeners are able to change their acoustic-to-phonetic mapping for fricative consonants based on the individual that they are listening to. However, the two groups make use of acoustic cues differently when performing this task. There was complete consistency among all NH listeners in relying more on VTL than any other cue, which makes sense because it is the physical property that should scale most directly with fricative resonance frequency. F0 was also used, but was relatively less influential in phonetic accommodation in NH listeners. Although F0 is an intuitive cue for direct identification of gender, and correlated with larger vocal tract sizes in everyday experience, F0 is controlled by an independent physical mechanism (the larynx) and also handled differently by the auditory system, which are possible explanations for the disparity in cue weights. Spectral contrast did not contribute in the hypothesized way.

Overall, it seems that the use of acoustic cues by CI listeners for various speech tasks is (1) different than that observed in NH listeners and (2) highly variable. It is reasonable to argue that if a particular cue is used consistently by the NH listeners, then it is the "correct" or "optimal" cue; CI users who used VTL might be considered to be better performers in the current task. Despite notoriously poor pitch perception among CI listeners in general, F0 is a more accessible cue than VTL, according to previous work by Fuller *et al.* (2014) and Gaudrain and Başkent (2018). It therefore was not surprising that F0 was revealed in the current study to be more influential than VTL as a cue for gender-related phonetic accommodation among CI listeners. However, it is also possible that the relative weighting of cues in this study was influenced by the vocal acoustics of the specific talkers involved. If these talkers were atypical in their relative disparity of VTL or F0, then the relative influence of these cues could emerge differently for other talkers. Furthermore, individuals might differ in the extent to which they express their gender identity through specific acoustic qualities of their voices.

Still unknown is what explains the individual variability among the CI listener group in terms of which cues are accessible and/or preferred in this perceptual task. None of the demographic factors from Table I (age, sex, device, etiology, CI experience, which ear) showed any clear relationship with the peak effect magnitude for the natural voices (i.e., the full acoustic switch from female to male voice), nor with any of the individual acoustic cues. It is possible that psychophysical measures of spectral resolution

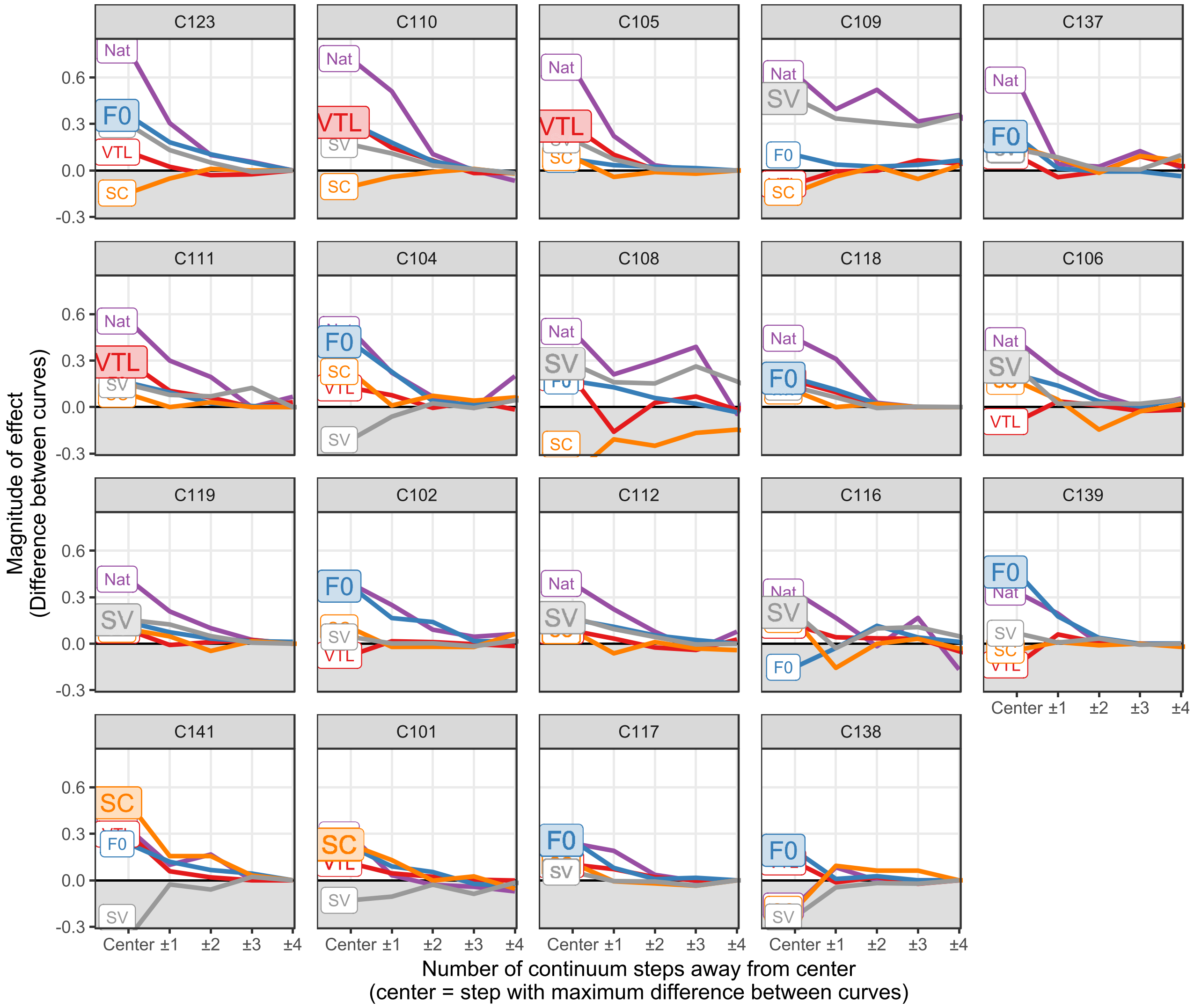

FIG. 8. (Color online) Influence of acoustic cues for phonetic accommodation for individual CI users. Term abbreviations are the same as for Fig. 7, with the leading cue for each listener highlighted with larger shaded text label. Listeners are arranged left to right, top to bottom, in order of the magnitude of the overall phonetic accommodation effect they show with the full complement of natural cues.

(e.g., eCAP slopes or psychophysical tuning curves) might shed light on the listener's ability to resolve formant structure, or that amplitude-modulation-based pitch ranking could explain the ability to use F0 cue. A previous study that measured both psychophysical auditory resolution and linguistic cue weighting (Winn *et al.*, 2016) showed some success in demonstrating correlations, but there was still substantial unexplained variability, likely because the prioritization of a cue within phonetic categorization does not necessarily follow from the listener's ability to access the cue.

Talker gender is a crucial aspect of a signal to hear, not only for the purpose of identifying a talker, but also because it aids in separating a target talker from background speech. Specifically, listeners with normal hearing (NH) can more easily hear a talker in the presence of a competing talker if the two talkers have different gender compared to if they are both women or both men (Brungart, 2001). CI listeners who are better at hearing acoustic cues for talker differences also tend to perform better in hearing speech in background noise as well (El Boghdady *et al.*, 2019). The auditory and psychological mechanisms underlying the differentiation of women's and men's voices are not fully understood, despite being a rather common part of everyday speech communication. Spectral contrast—the hypothesized mechanism that would have accorded with biologically framed theories of speech perception (Kluender *et al.*, 2003; Sjerps and Reinisch 2015; Winn and Stilp, 2019)—turned out to not explain the phonetic accommodation behavior, leaving more questions for future work.

As mentioned in the Introduction, there is a complex interdependence of F0 and VTL information as listeners identify talkers and vowels. For languages that do not incorporate F0 to determine lexical tone, F0 can be estimated rather independently of the identity of the speech sound itself. In contrast, formant frequencies change largely due to the identity of the vowel or consonant being spoken, which should complicate their use as a paralinguistic cue. Using formants for talker identification therefore should require awareness of the target phoneme being spoken. Corroborating this, vowel identification is related to accuracy of judging talker sex (Eklund and Traunmüller, 1997). Barreda and Nearey (2012, 2013) further suggest that the F0 cue can be used by listeners to infer talker characteristics that would contextualize or normalize the vowel formant information (as opposed to F0 playing a role in vowel identification directly). For example, given a low F0, the listener might assume a male speaker and thus expect lower formant frequencies [akin to the hypothesized "proxy" cue of a talker's face influencing gender-related phoneme categorization observed by Johnson *et al.* (1999) and Winn *et al.* (2013)]. This type of F0 perception and transfer is likely an inaccessible skill for CI listeners, since the representation of F0 is primarily in the periodicity of the envelope, which is a weak cue compared to the harmonic pitch obtained in normal acoustic hearing. Furthermore, electrode activation diagrams published by Gaudrain and Başkent (2018) suggest that the difference between stereotypically male and female voices might not be a matter of lower versus higher-rate periodicity, but rather the presence versus absence of periodicity, since a woman's voice might have a repetition rate too fast to be cleanly represented by the envelope of the electrode activation. The results of the current study are not in conflict with the results of Barreda and Nearey, but diverge in method, since the current study explicitly treated VTL (spectral shape) and F0 (periodicity rate) independently, since they are handled differently by the auditory system, and also investigated the effects of vowel acoustics on perception of adjacent consonants rather than on perception of the vowels themselves.

With the current results in mind, it is reasonable to suspect that previous studies that identified gender-related phonetic accommodation might have involved two non-exclusive types of accommodation at once. First, there is clearly adjustment for acoustic parameters corresponding to vocal tract size, as the VTL cue was very influential in the current study. These perceptual patterns might arise out of talker-size perception independently of gender. The fact that the Winn *et al.* (2013) study showed different categorization functions for each of the two women and two men in the study further suggests that a binary perception of gender does not explain all of the results. However, there is also a second mechanism which could facilitate perceptual adjustments based on expectations related to gendered speech, even in the absence of VTL differences. Evidence for such a mechanism is found in studies where visual cues to gender are coupled with androgynous voices (Johnson *et al.*, 1999) or voices whose acoustic cues to gender are impoverished via the use of a CI or by noise vocoding (Winn *et al.*, 2013). These previous studies arguably suggest that visual cues can serve as "proxy" cues for vocal tract size (just as F0 might be considered a proxy cue because of its common covariation with VTL), or the visual cues might instead suggest gender stereotypicality independent of VTL, which could be intentionally expressed as differences in phoneme articulation. Whereas CI listeners might infer these proxy cues on account of not having clear access to VTL in the acoustic/electric signal, the NH listeners are less likely to rely on proxy cues because of their precise coding of the frequency cues for VTL without any need to gather indirect cues.

V. CONCLUSIONS

When perceptually adjusting to gender-related differences in voice acoustics, listeners with normal hearing primarily rely on acoustic cues that correspond to VTL, with less reliance on voice pitch. The difference between cues used for gender-related phonetic accommodation and direct gender identification suggest that these are two distinct processes; perhaps previous accounts of gender accommodation are consistent with talker-size estimation rather than gender perception *per se*. Listeners with cochlear implants demonstrate gender-related phonetic accommodation, but show substantial individual variability in their weighting of acoustic cues, with F0 (pitch) emerging as the strongest cue on average, consistent with earlier work examining the relative advantage of perceiving F0 versus VTL with electric hearing.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health Grant No. NIH NIDCD R01 DC017114 (Winn). Data collection and participant recruitment were assisted by Kate Teece, Heather Kreft, Moira McShane, Tiffany Mitchell, Josephine Lyou, Siuho Gong, Steven Gianakas, Emily Hugo, Paula Rodriguez, Hannah Matthys, and Lindsay Williams.

[1] See supplementary material at https://doi.org/10.1121/10.0001672 for detailed parameters of the stimulus continuum, and individual rankings of each context effect using both logit and proportional models.

Akaike, H. (**1974**). "A new look at the statistical model identification," IEEE Trans. Auto. Control **19**, 716–723.

Andreeva, B., Demenko, G., Wolska, M., Möbius, B., Zimmerer, F., Jügler, J., Oleskowicz-Popiel, M., and Trouvain, J. (**2014**). "Comparison of pitch range and pitch variation in Slavic and Germanic languages," in *Proceedings of the 7th International Conference on Speech Prosody*, edited by N. Campell, D. Gibson, and D. Hirst, pp. 776–780.

Barreda, S. (**2016**). "Investigating the use of formant frequencies in listener judgments of talker size," J. Phon. **55**, 1–18.

Barreda, S. (**2017**). "An investigation of the systematic use of spectral information in the determination of apparent-talker height," J. Acoust. Soc. Am. **141**, 4781–4792.

Barreda, S., and Nearey, T. (**2012**). "The direct and indirect roles of fundamental frequency in vowel perception," J. Acoust. Soc. Am. **131**, 466–477.

Barreda, S., and Nearey, T. (**2013**). "Training listeners to report the acoustic correlate of formant-frequency scaling using synthetic voices," J. Acoust. Soc. Am. **133**, 1065–1077.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**, 1–48.

Boersma, P., and Weenink D. (**2017**). "Praat: Doing phonetics by computer" [computer program], version 6.0.26, http://www.fon.hum.uva.nl/praat/ (Last viewed March 14, 2017).

Boëx, C., de Balthasar, C., Kós, M., and Pelizzone, M. (**2003**). "Electrical field interactions in different cochlear implant systems," J. Acoust. Soc. Am. **114**, 2049–2057.

Brungart, D. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Chatterjee, M., and Shannon, R. (**1998**). "Forward masked excitation patterns in multielectrode electrical stimulation," J. Acoust. Soc. Am. **103**, 2565–2572.

Chodroff, E., and Wilson, C. (**2020**). "Acoustic-phonetic and auditory mechanisms of adaptation in the perception of sibilant fricatives," Atten. Percept. Psychophys. **82**, 2027–2048.

DeVries, L., Scheperle, R., and Bierer, J. A. (**2016**). "Assessing the electrode-neuron interface with the electrically evoked compound action potential, electrode position, and behavioral thresholds," J. Assoc. Res. Otolaryngol. **17**, 237–252.

Eklund, I., and Traunmüller, H. (**1997**). "Comparative study of male and female whispered and phonated versions of the long vowels of Swedish," Phonetica **54**, 1–21.

El Boghdady, N., Gaudrain, E., and Başkent, D. (**2019**). "Does good perception of vocal characteristics relate to better speech-on-speech intelligibility for cochlear implant users?," J. Acoust. Soc. Am. **145**, 417–439.

Fant, G. (**1966**). "A note on vocal tract size factors and non-uniform F-pattern scalings," Prog. Stat. Rep. Comput. Ling. **29**, 22–30.

Fu, Q.-J., Chinchilla, S., Nogaki, G., and Galvin, J. J. III (**2005**). "Voice gender identification by cochlear implant users: The role of spectral and temporal resolution," J. Acoust. Soc. Am. **118**, 1711–1718.

Fuller, C., Gaudrain, E., Clarke, J., Galvin, J., Fu, Q.-J., Free, R., and Başkent, D. (**2014**). "Gender categorization is abnormal in cochlear implant users," J. Res. Otolaryng. **15**, 1037–1048.

Gaudrain, E., and Başkent, D. (**2015**). "Factors limiting vocal-tract length discrimination in cochlear implant simulations," J. Acoust. Soc. Am. **137**, 1298–1308.

Gaudrain, E., and Başkent, D. (**2018**). "Discrimination of voice pitch and vocal-tract length in cochlear implant users," Ear Hear. **39**, 226–237.

Hedrick, M., and Carney, A. (**1997**). "Effect of relative amplitude and formant transitions on perception of place of articulation by adult listeners with cochlear implants," J. Speech Lang. Hear. Res. **40**, 1445–1457.

Hillenbrand, J., and Clark, M. (**2009**). "The role of f0 and formant frequencies in distinguishing the voices of men and women," Atten. Percept. Psychophys. **71**, 1150–1166.

Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Holden, L., Finley, C., Firszt, J., Holden, T., Brenner, C., Potts, L., Gotter, B., Vanderhoof, S., Mispagel, K., Heydebrand, G., and Skinner, M. (**2013**). "Factors affecting open-set word recognition in adults with cochlear implants," Ear Hear. **34**, 342–360.

Jaekel, N. B., Newman, R., and Goupell, M. (**2017**). "Speech rate normalization and phonemic boundary perception in cochlear-implant users," J. Speech Lang. Hear. Res. **60**, 1398–1416.

Johnson, K., Strand, E., and D'Imperio, M. (**1999**). "Auditory-visual integration of talker gender in vowel perception," J. Phon. **27**, 359–384.

Jongman, A., Wayland, R., and Wong, S. (**2000**). "Acoustic properties of English fricatives," J. Acoust. Soc. Am. **108**, 1252–1263.

Kluender, K., Coady, J., and Kiefte, M. (**2003**). "Sensitivity to change in perception of speech," Speech Commun. **41**, 59–69.

Kovačić, D., and Balaban, E. (**2009**). "Voice gender perception by cochlear implantees," J. Acoust. Soc. Am. **126**, 762–775.

Landsberger, D., Svrakic, J., and Svirsky, M. (**2015**). "The relationship between insertion angles, default frequency allocations, and spiral ganglion place pitch in cochlear implants," Ear Hear. **36**, e207.

Liberman, M. (**2013**). "Biology, sex, culture, and pitch," Blog post on *Language Log*, dated August 16, 2013, https://languagelog.ldc.upenn.edu/nll/?p=5908 (Last viewed January 2, 2020).

Lotto, A., and Kluender, K. (**1998**). "General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification," Percept. Psychophys. **60**, 602–619.

Luo, X., and Fu, Q.-J. (**2006**). "Contribution of low-frequency acoustic information to Chinese speech recognition in cochlear implant simulations," J. Acoust. Soc. Am. **120**, 2260–2266.

Mann, V., and Repp, B. (**1980**). "Influence of vocalic context on perception of the /ʃ/-/s/ distinction," Percept. Psychophys. **28**, 213–228.

Maryn, Y., Roy, N., De Bodt, M., Van Cauwenberge, P., and Corthals, P (**2009**). "Acoustic measurement of overall voice quality: A meta-analysis," J. Acoust. Soc. Am. **126**, 2619–2634.

McMurray, B., and Jongman, A. (**2011**). "What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations," Psych. Rev. **118**, 219–246.

Miller, J. (**1981**). "Some effects of speaking rate on phonetic perception," Phonetica **38**, 159–180.

Moberly, A., Lowenstein, J., and Nittrouer, S. (**2016**). "Word recognition variability with cochlear implants: 'Perceptual attention' versus 'auditory sensitivity,' " Ear Hear. **37**, 14–26.

Munson, B. (**2011**). "The influence of actual and imputed talker gender on fricative perception, revisited," J. Acoust. Soc. Am. **130**, 2631–2634.

Munson, B., Donaldson, G., Allen, S., Collison, E., and Nelson, D. (**2003**). "Patterns of phoneme misperceptions by individuals with cochlear implants," J. Acoust. Soc. Am. **113**, 925–935.

Munson, B., Jefferson, S., and McDonald, E. (**2006**). "The influence of perceived sexual orientation on fricative identification," J. Acoust. Soc. Am. **119**, 2427–2437.

R Core Development Team (**2016**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, version 3.3.2 https://www.R-project.org/ (Last viewed 7/25/2020).

Shannon, R., Fu, Q.-J., and Galvin, J. (**2004**). "The number of spectral channels required for speech recognition depends on the difficulty of the listening situation," Acta Otolargol. **552**, 50–54.

Sjerps, M., and Reinisch, E. (**2015**). "Divide and conquer: How perceptual contrast sensitivity and perceptual learning cooperate in reducing input variation in speech perception," J. Exp. Psych. Human Percept. Perform. **41**, 710–722.

Skuk, V., and Schweinberger, S. (**2014**). "Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice and gender," J. Speech Lang. Hear. Res. **57**, 285–296.

Stilp, C. (**2020**). "Acoustic context effects in speech perception," in *Wiley Interdisciplinary Reviews: Cognitive Science* Vol. **11**, pp. 1–18.

Stilp, C., Anderson, P., Assgari, A., Ellis, G., and Zahorik, P. (**2016**). "Speech perception adjusts to stable spectrotemporal properties of the listening environment," Hear Res. **341**, 168–178.

Stilp, C., Anderson, P., and Winn, M. (**2015**). "Predicting contrast effects following reliable spectral properties in speech perception," J. Acoust. Soc. Am. **137**, 3466–3476.

van Dommelen, W., and Moxness, B. (**1995**). "Acoustic parameters in speaker height and weight identification: Sex-specific behaviour," Lang. Speech **38**, 267–287.

Winn, M. (**2020**). "Accommodation of gender-related phonetic differences by listeners with cochlear implants and in a variety of vocoder simulations," J. Acoust. Soc. Am. **147**, 174–190.

Winn, M., Chatterjee, M., and Idsardi, W. (**2012**). "The use of acoustic cues for phonetic identification: Effects of spectral degradation and electric hearing," J. Acoust. Soc. Am. **131**, 1465–1479.

Winn, M., and Litovsky, R. (**2015**). "Using speech sounds to test functional spectral resolution in listeners with cochlear implants," J. Acoust. Soc. Am. **137**, 1430–1442.

Winn, M., Rhone, A., Chatterjee, M., and Idsardi, W. (**2013**). "Auditory and visual context effects in phonetic perception by normal-hearing listeners and listeners with cochlear implants," Front. Psych: Aud. Cogn. Neurosci. **4**(824), 1–13.

Winn, M., Won, J. H., and Moon, I. J. (**2016**). "Assessment of spectral and temporal resolution in cochlear implant users using psychoacoustic discrimination and speech cue categorization," Ear Hear. **37**(6), e377–e390.

Winn, M., and Stilp, C. (**2019**). "Phonetics and the auditory system," in *The Routledge Handbook of Phonetics*, edited by W. Katz and P. Assmann (Routledge, New York), pp. 164–192.

Zeng, F.-G., Grant, G., Niparko, J., Galvin, J., Shannon, R., Opie, J., and Segel, P. (**2002**). "Speech dynamic range and its effect on cochlear implant performance," J. Acoust. Soc. Am. **111**, 377–386.