JASA ARTICLE

Accommodation of gender-related phonetic differences by listeners with cochlear implants and in a variety of vocoder simulations

Matthew B. Winn[a)]
Department of Speech & Hearing Sciences, University of Minnesota, 164 Pillsbury Drive Southeast, Minneapolis, Minnesota 55455, USA

ABSTRACT:
Speech perception requires accommodation of a wide range of acoustic variability across talkers. A classic example is the perception of "sh" and "s" fricative sounds, which are categorized according to spectral details of the consonant itself, and also by the context of the voice producing it. Because women's and men's voices occupy different frequency ranges, a listener is required to make a corresponding adjustment of acoustic-phonetic category space for these phonemes when hearing different talkers. This pattern is commonplace in everyday speech communication, and yet might not be captured in accuracy scores for whole words, especially when word lists are spoken by a single talker. Phonetic accommodation for fricatives "s" and "sh" was measured in 20 cochlear implant (CI) users and in a variety of vocoder simulations, including those with noise carriers with and without peak picking, simulated spread of excitation, and pulsatile carriers. CI listeners showed strong phonetic accommodation as a group. Each vocoder produced phonetic accommodation except the 8-channel noise vocoder, despite its historically good match with CI users in word intelligibility. Phonetic accommodation is largely independent of linguistic factors and thus might offer information complementary to speech intelligibility tests which are partially affected by language processing. https://doi.org/10.1121/10.0000566

(Received 30 April 2019; revised 6 December 2019; accepted 13 December 2019; published online 22 January 2020)
[Editor: Deniz Baskent] Pages: 174–190

I. INTRODUCTION

Speech intelligibility is a complicated process that can be described and evaluated at many levels. Accuracy scores for words and sentences are routine in audiological evaluations and are commonly obtained for speech spoken by a single talker at a time. However, in everyday communication speech perception requires accommodation of a wide range of acoustic differences between talkers. In addition to the measurement of overall word intelligibility, it is also desirable to measure the extent to which listeners can accommodate this type of talker-related variability in acoustics. The current study follows previous work (Mann and Repp, 1980; Johnson *et al.*, 1999; Munson *et al.*, 2006; Winn *et al.*, 2013) that used perceptual tests designed to probe phonetic accommodation of gender by explicitly modeling boundaries between phonemes that are affected by talker variability. The goal of the current study is to explore the extent to which acoustic degradation affects how listeners adjust for acoustic differences related to talker gender as they are perceiving phonemes—an ability not necessarily revealed by word-level accuracy scores.

One relatively well-studied example of inter-talker phonetic variability is the difference in the production of fricatives /ʃ/ ("she") and /s/ ("see"). These sounds are perceived as fricatives of relatively lower and higher frequency, respectively, although the acoustic properties can be complex (McMurray and Jongman, 2011). When the vocal tract resonance frequencies are globally lower—as for a man's voice—the perceived boundary (in frequency space) between the low /ʃ/ and high /s/ is accordingly shifted to a lower frequency than for a woman's voice (Mann and Repp, 1980; Jongman *et al.*, 2000). This is an example of phonetic accommodation of talker gender, or more generally, a phonetic context effect. Ambiguous fricatives tend to be labeled /s/ if spliced onto the voice (vowel) of a talker perceived to be man, and /ʃ/ if spliced onto the voice of someone perceived to be a woman. Thus, the tendency to adjust fricative categorization in this manner reflects one's ability to adjust to gender-related talker acoustics—particularly in the *spectral* domain, as the fricatives do not have any appreciable differences in the temporal domain. This phonetic accommodation behavior serves as the primary outcome measure in the current study.

Word- and sentence-repetition accuracy scores for degraded speech are thought to be driven largely by auditory processing but is also at least partly affected by linguistic processing at the level of sentences (Patro and Mendel, 2016; Winn, 2016) and individual words (Gianakas and Winn, 2019). Even at the level of individual consonants and vowels presented without any linguistic context, listeners might be biased to report whichever phonetic category is most appropriate for their native language, even if the input was not a perfect auditory match. Especially among people

a)Electronic mail: mwinn@umn.edu

with hearing impairment, there is wide variation in the ability to compensate for missing information, including ways that do not show up in intelligibility scores (Başkent *et al.*, 2016). It would, therefore, be beneficial to complement intelligibility tests with auditory tasks that are sensitive to acoustic details that are not subject to a linguistic influence, and yet are still relevant to everyday communication. This study attempts to demonstrate such a test by measuring accommodation of phonetic differences across a wide variety of listening conditions, where the accommodation would not change the identity of the signal so much that it should be recognized incorrectly, yet still be detectable in behavioral tasks that probe phonetic perception in a granular fashion.

It is known that listeners with normal hearing (NH) and with cochlear implants (CIs, described further below) can both demonstrate phonetic accommodation of gender-related variability in talker acoustics when categorizing fricative sounds. The current study extends an earlier study of a small number of CI listeners (Winn *et al.*, 2013) and explores the issue of testing phonetic accommodation with spectrally degraded signals that could be used to simulate cochlear implants in listeners who have typical acoustic hearing.

A. CIs

CIs are auditory neural prostheses that restore a sense of hearing to people who have severe to profound deafness. Although they have been remarkably successful as a treatment for hearing loss, there are still many aspects of CIs that remain incompletely understood, such as how to evaluate their ability to transmit speech information. The intuitive way to measure this is to use speech intelligibility tests, but as discussed above, there are some challenges inherent in using intelligibility scores. It would be useful to have some auditory tasks for CI users that probe transmission of speech information that is not subject to such influences and yet still is relevant to everyday speech communication.

There are some standardized approaches to avoid linguistic influence when testing auditory perception, including using broadband rippled spectra (Won *et al.*, 2007) or spectro-temporally modulated rippled spectra (Aronoff and Landsberger, 2013). However, such stimuli have not been found to have any meaningful acoustic correspondence to speech in either the spectral or the temporal domains, and therefore it is difficult to explain mechanisms of how they could explain speech perception abilities. The current study probes perception of acoustic details contained with the speech stimulus itself in an attempt to ensure ecological relevance, albeit without any ability to quantify spectral or temporal resolution directly.

B. Simulating cochlear implants with vocoders

One of the main tools that is used to understand perception of speech with CIs is the vocoder, which has been instrumental in demonstrating the robustness of speech perception (Shannon *et al.*, 1995) and in predicting intelligibility among those who use CIs (cf. Friesen *et al.*, 2001). A vocoder divides the frequency spectrum into a number of discrete bands (typically equally distributed according to simulated cochlear spacing) and represents the amplitude envelope of each of those bands using a simplified carrier, such as a pure tone or narrowband noise that is matched (or at least monotonically related) to its corresponding original input frequency band. This is a common mechanism in modern CIs as well; these devices typically transmit the envelope of discrete frequency bands using pulsatile stimulation that discards temporal fine structure, as does a vocoder. In this sense, the vocoder is an implementation of one of the fundamental aspects of CI processing and sound transmission.

A vocoder allows an experimenter to parametrically vary specific processing parameters across listeners without the uncontrollable variability related to real CI recipients. It is thought that better-performing CI listeners will achieve speech intelligibility scores that closely match scores obtained with listeners with NH who are responding to speech that is vocoded with six to eight channels of spectral information, regardless of carrier type (Dorman *et al.*, 1997; Friesen *et al.*, 2001). Eight-channel vocoders or noise vocoders, in general, have therefore sometimes been referred to simply as "CI simulations" (Shannon *et al.*, 2004; Chatterjee and Peng, 2008; Pals *et al.*, 2013; Aronoff *et al.*, 2015).

Apart from manipulating just the number of vocoder carrier channels, experimenters have also explored manipulation of properties like envelope fidelity (Xu *et al.*, 2005), dynamic range (Stafford *et al.*, 2014), frequency shifting (Rosen *et al.*, 1999; Zhou *et al.*, 2010), frequency compression (Başkent and Shannon, 2005), spread of excitation (Litvak *et al.*, 2007; Winn and Litovsky, 2015; Grange *et al.*, 2017), interaction between frequency analysis channels (Crew *et al.*, 2012; Oxenham and Kreft, 2014, 2016), and the distribution of frequency energy around simulated cochlear dead regions (DiNino *et al.*, 2016). Thus, it is clear that vocoders can be used to explore a large number of important factors relating to CIs, each of which might require different kinds of stimuli to differentiate contributions among parameters of interest. Laneau *et al.* (2006) specifically pointed this out in a study where vocoder parameters needed to match CI speech intelligibility were different than the parameters needed to match CI pitch perception. The current study uses a variety of vocoders that could potentially be used to simulate listening with a CI, to see if the parameters have demonstrable effects on phonetic accommodation specifically.

C. Previous studies of phonetic accommodation

In a study by Winn *et al.* (2013), CI listeners demonstrated not only good identification of natural-sounding /s/ and /ʃ/ sounds but also near-normal phonetic accommodation of talker gender (which was further enhanced by

accompanying complementary visual cues). This was surprising considering the voice-related cues that drive the phonetic context effect are primarily spectral in nature, and CI users are known to have an especially poor spectral resolution. It was also surprising because NH participants who heard sounds processed with an 8-channel noise vocoder showed virtually no phonetic accommodation effect at all. It was not clear whether the lack of accommodation with the 8-channel noise vocoder was attributable to the lack of experience of NH participants with degraded speech signals, whether that particular vocoder simply was not a good approximation of the signal in the cochlear implant or some combination of these and other factors.

D. Acoustic cues for talker gender

Adjusting acoustic criteria for phoneme identification based on talker gender is not the same as directly identifying talker gender, but it is worth considering the acoustics of gender and the ability of CI listeners to identify talker gender directly. Numerous acoustic cues are available to distinguish the voices of women and men, including fundamental frequency, formant frequencies (henceforth vocal-tract length), and spectrum level (Skuk and Schweinberger, 2014). Studies by Fu *et al.* (2004) and Kovačić and Balaban (2009) suggest that fundamental frequency (*F*0) plays a role in gender identification by CI listeners, as performance was worse in both studies when the disparity in *F*0 between women and men was smaller. Fuller *et al.* (2014) further clarified this issue by demonstrating relatively higher perceptual weighting of *F*0 cues compared to orthogonal vocal-tract length (VTL) cues in a group of 19 CI listeners who identified talker gender. One might expect that *F*0 perception would be limited in CI listeners as there is no clear harmonic structure and only relatively weak rate-pitch cues in the temporal envelope. VTL is also difficult for CI listeners to ascertain (Gaudrain and Başkent, 2018) consistent with formant structure being generally more difficult to perceive by that group (Winn *et al.*, 2012; Winn and Litovsky, 2015). 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.

E. New questions and hypotheses in the current study

In light of the disagreement between the aforementioned CI results and 8-channel noise vocoder results, the goal in the current study was to use a phonetic accommodation test to ask four new questions: (1) Is experience with degraded signals necessary to show a phonetic accommodation effect? (i.e., did the vocoder results by Winn *et al.*, 2013 not demonstrate a phonetic accommodation effect because of the acute listening conditions?) (2) Do different kinds of vocoders lead to a better match to CI performance? (3) Do systematic changes in spectral resolution result in corresponding changes in phonetic accommodation, and (4) Are these outcome measures any more differentiable than word-recognition scores in the same vocoder conditions?

The hypotheses were (1) experience with degraded signals would not be necessary, as pilot testing revealed that vocoder resolution intermediate to the 8-channel vocoder and normal speech (i.e., a vocoder with greater than eight channels) was sufficient to elicit a gender-related accommodation effect, (2) vocoders that better match the frequency-channel allocation of the CI listeners will elicit results that are more similar to CI listeners, (3) vocoders with better spectral resolution should grant better perception of talker properties and therefore elicit a more substantial talker context effect, and (4) the results of the phonetic accommodation test will be qualitatively more differentiable than word-recognition results, but difficult to directly compare because the results are in different domains.

II. METHODS

A. Participants

There were 20 listeners with cochlear implants, including seven unilateral recipients from the Winn *et al.* (2013) study conducted at the University of Maryland, nine bilateral recipients recruited at the University of Wisconsin-Madison, and four (one bilateral, three unilateral) recruited at the University of Washington in Seattle. All CI participants were between the ages of 50 and 84. They each were tested using their normal processor configurations (unilateral or bilateral), with their most commonly used program. Three CI listeners who also wore hearing aids were instructed to remove the hearing aids prior to testing. There were also 48 listeners with NH, defined as having pure-tone thresholds <20 dB hearing level (HL) from 250 to 8000 Hz in both ears, of whom 14 were recruited at the University of Wisconsin-Madison, and 34 at the University of Washington.

The NH listeners had an age range from 18 to 38, with a median age of 23 years. The larger number of NH listeners were used to test a variety of vocoder conditions to be described below. All participants gave informed consent that was approved at the respective institutions in which the data were collected. All participants were native speakers of American English except for two NH listeners who spoke fluent English, and whose native language contained a /s/–/ʃ/ contrast. One listener was excluded because her native language (Tamil) did not contain this contrast in all phonetic environments.

B. Stimuli for phonetic accommodation test

1. Overview

Stimuli for the main speech categorization test were monosyllabic words that sounded like "sue," "see," "shoe," and "she." These stimuli can be understood as having two major components: the *fricative* itself (the "target" phoneme), and the *context*, which is the vowel segment that was appended to the fricative. The current experiment used

https://doi.org/10.1121/10.0000566

a subset of the stimuli used by Winn *et al.* (2013), which included four parameters: fricative spectrum, talker gender, vowel (/i/ or /u/), and formant transitions within the vowel that were from an original /s/-onset or /ʃ/-onset syllable. In the current study, vowels were excised from /s/-onset words because the formant transition effect is rather small, and the /s/-onset vowels yield natural-sounding syllables regardless of which fricative is added.

It is important to note that the value of the current task does not hinge on any special importance of /ʃ/ and /s/ sounds in spoken language. Rather, it is that this pair of sounds provides an environment that can be used to probe a listener's ability to extract information that is relevant to the task of accommodating talker differences. It should, therefore, in theory, be a test that probes for an auditory skill that is more subtle than word intelligibility and not affected by linguistic knowledge. It is also suitable as a test across languages since the /ʃ/–/s/ contrast is expressed in a large majority of the world's languages. The current test might, therefore, hold promise as a new complementary measure that could be used to test speech signal transmission without an undesirable influence of linguistic-cognitive processing.

2. Fricative synthesis

The fricative components of the stimuli gradually morphed from /ʃ/ to /s/ along multiple frequency dimensions. A nine-step fricative continuum had endpoints that were modeled after productions of /ʃ/ and /s/ sounds in real words ("see, she, sue, shoe") produced by women and men recorded for the Winn *et al.* (2013) study.

Stimuli were synthesized by combining filtered bands of noise. The filtered noises were created using the Praat software (Boersma and Weenink, 2011) using a procedure illustrated in Fig. 1. White noise of 180 ms duration was filtered into three peaks of specified frequency, bandwidth and amplitude. Filtering was done using Hann bands that operated in the frequency domain, creating filter slopes that ranged from −9 to −12 dB/octave across the continuum. The narrowband frequency peaks were summed and then shaped with a uniform amplitude contour with 115 ms rise-time and 18 ms fall-time, which were representative of these consonants across the recordings that were collected in the original study. Spectral peak frequencies were interpolated along a log-frequency scale. Table I contains details of the parameters of this continuum, which are illustrated in Fig. 2.

3. Contexts

Each step of the fricative continuum was prepended to each of eight vocalic contexts consisting of the /i/ and /u/ vowels from natural recordings from the words "see" and "sue" spoken by four native speakers of English (two female and two male, all phonetically trained, one of whom was the author).

C. Monosyllabic words

To validate the stimulus processing in the current study against the common standard measure of word recognition, a separate group of 14 NH listeners was tested using monosyllabic words (e.g., "boat," "dime," "take," "run") in each of the vocoder conditions. The words were drawn from the Maryland CNC corpus (Peterson and Lehiste, 1962), which was designed to contain words that are familiar to most listeners and yet difficult enough to differentiate among

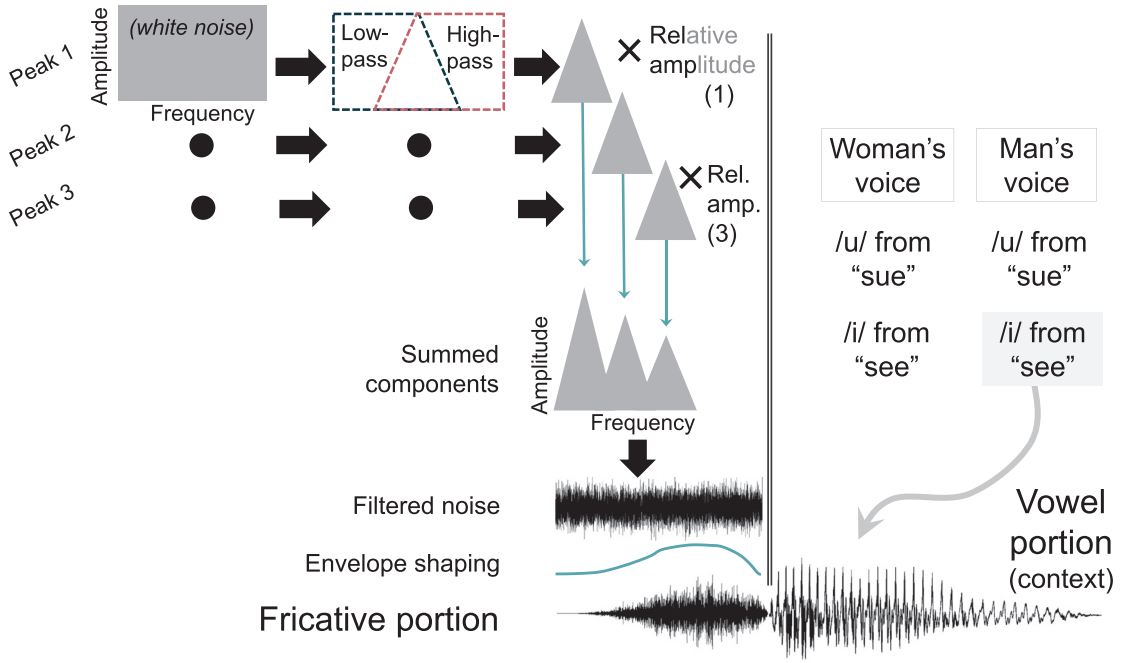

FIG. 1. Construction of the stimuli for the phonetic accommodation test. The fricatives were created by summing three filtered noises and then preappended to a vowel drawn from one of four vocalic contexts listed on the right side of the figure. Spectrum peaks are not drawn to scale in this image.

TABLE I. Acoustic description of the fricative continuum. Note: the peak frequency continuum steps were interpolated using a log scale ranging from 3.467 to 3.799 but presented here in Hz. SP1, SP2, SP3 refer to the three spectral peaks from lowest to highest.

Continuum step:	/ʃ/ 1	2	3	4	5	6	7	8	9 /s/
Peak frequencies (Hz)									
SP1	2932	3226	3550	3906	4298	4729	5203	5726	6300
SP2	6130	6357	6592	6837	7090	7352	7625	7907	8200
SP3	8100	8283	8472	8666	8863	9065	9272	9484	9700
Amplitude relative to peak 2 (dB)									
SP1	1.67	0.83	0.00	−0.83	−1.67	−2.50	−3.33	−4.17	−5
SP3	−1.7	−0.8	0.0	0.8	1.7	2.5	3.3	4.2	5

various degrees of hearing impairment. The words were presented in isolation in quiet, with no carrier phrase.

D. Vocoders

There were three types of vocoders used for this study, and each was implemented with two different levels of spectral resolution.

1. Continuous-interleaved-sampling ("CIS")-style channel noise vocoder

The first type of vocoder replicated continuous-interleaved-sampling (CIS) style processing. It was a discrete-channel vocoder that divided the spectrum into frequency bands and represented those bands with rectangular bands of noise. This vocoder, inspired by the one used by Shannon *et al.* (1995), was implemented with 8 or 24 channels, reflecting poorer and better spectral resolution, respectively. The 8-channel iteration was chosen because it is generally thought to be an adequate model for CI perception and because 8 channels were used in the prior study by Winn *et al.* (2013). The 24-channel vocoder was chosen so that there would be a condition with the same principles of signal processing but with a better spectral resolution, to test the hypothesis (#3) that this factor would change the outcome measure.

The frequency channels were calculated to have equal cochlear spacing according to the function published by Greenwood (1990), using a 35 mm cochlear length. Bands were extracted from the speech signals using Hann filters in Praat software, which operate linearly in the frequency domain. Each filter was 50 Hz wide, meaning that the gain of a high-pass filter was zero at 25 Hz below the frequency boundary, and 100% at 25 Hz above the frequency boundary, with intermediate values that were linearly interpolated. The 50 Hz bandwidth was used to prevent ringing artifacts that would be present with steeper filters. Specific band frequency cutoff values are listed in Table II. The amplitude envelope for each band was used to modulate a band of white noise that was then filtered to have the same bandwidth as the original band of speech. The envelope in each speech band was extracted using the IntensityTier function in the Praat software, and low-pass filtered with a 300-Hz cutoff frequency. The modulated and filtered bands of noise were summed to create the final vocoded version of each speech signal.

2. Advanced combination encoder ("ACE")-style noise vocoder

The second type of vocoder also used noise-band carriers, but with frequency-channel allocation and channel peak-picking modeled after the ACE stimulation strategy commonly used in the "Nucleus" family of implants developed by the Cochlear corporation. For this study, this vocoder will thus be referred to as the "ACE" vocoder. The channel-frequency allocation was obtained through the clinical fitting (mapping) software for the Cochlear device; no consultation or support was provided by the device manufacturer. Filtering and envelope extraction for the channel analysis were done using Hann filters as described for the previous vocoder. After creating all 22 channels using this

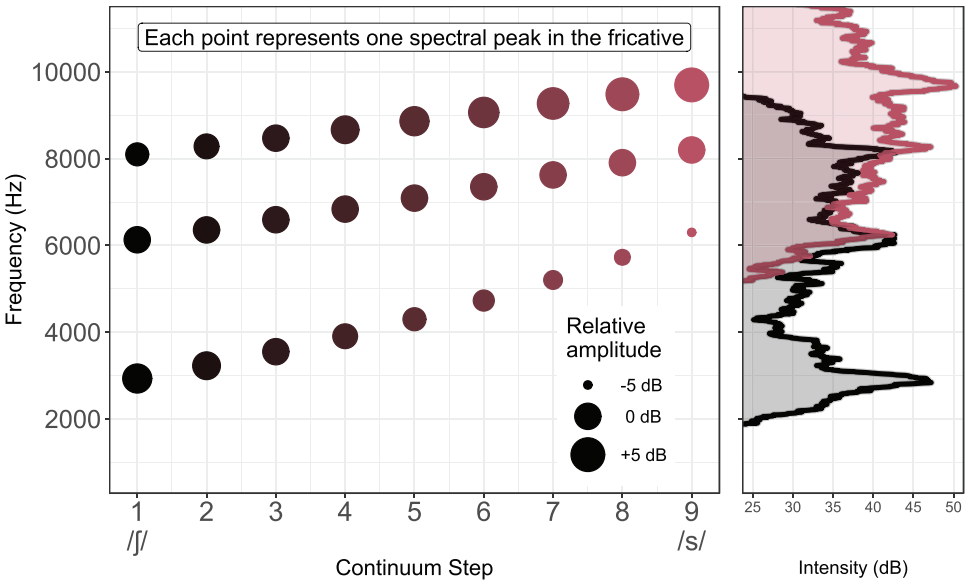

FIG. 2. Parameters of the fricative continuum that was labeled in the perceptual experiment. Across each step, there were three spectral peaks (shown as three dots) defined by their central frequency and relative amplitude (normalized to the central peak), shown as the relative size of the dots. On the right margin, spectra of the continuum endpoints aligned with the parameters indicated in the main figure panel.

https://doi.org/10.1121/10.0000566

TABLE II. Analysis and carrier filter band corner frequencies for the vocoders.

	8 Channel			24 Channel			ACE			PTP (pulsatile)		
Chan #	Low	Center	High	Low	Center	High	Low	Center	High	Low	Center	High
1	150	218	302	150	171	194	188	250	313	238	329	442
2	302	405	532	194	218	244	313	375	438	442	506	578
3	532	688	879	244	272	302	438	500	563	578	611	646
4	879	1116	1406	302	334	368	563	625	688	646	711	782
5	1406	1763	2203	368	405	444	688	750	813	782	879	986
6	2203	2744	3410	444	486	532	813	875	938	986	1083	1189
7	3410	4228	5236	532	580	632	938	1000	1063	1189	1287	1393
8	5236	6475	8000	632	688	747	1063	1125	1188	1393	1523	1665
9				747	811	879	1188	1250	1313	1665	1828	2005
10				879	953	1031	1313	1425	1563	2005	2200	2413
11				1031	1116	1206	1563	1650	1813	2413	2641	2889
12				1206	1302	1406	1813	1925	2063	2889	3180	3500
13				1406	1517	1636	2063	2175	2313	3500	3825	4180
14				1636	1763	1900	2313	2500	2688	4180	5810	8054
15				1900	2046	2203	2688	2875	3063			
16				2203	2371	2551	3063	3300	3563			
17				2551	2744	2951	3563	3800	4063			
18				2951	3172	3410	4063	4350	4688			
19				3410	3664	3936	4688	5000	5313			
20				3936	4228	4541	5313	5675	6063			
21				4541	4876	5236	6063	6500	6938			
22				5236	5621	6033	6938	7500	7938			
23				6033	6475	6949						
24				6949	7456	8000						

method, a series of overlapping 30-ms time windows (with 15 ms overlap) were analyzed to determine which channels were picked or dropped. This window size was chosen not to replicate any aspect of CI processing, but instead to preserve dynamic changes that would be sufficient for transmission of various phonetic cues relating to place of articulation and fricatives (Blumstein and Stevens, 1979; Jongman, 1989; Assmann, 1996). This 30 ms duration can be regarded as a window for updating the spectral envelope for peak-picking analysis rather than a sampling rate for the intensity of the signal. The eight channels with the highest amplitude after pre-emphasis were maintained in the carrier signal for each time bin, and other channels were dropped by setting the envelope to zero. Within each time window, the envelope was still sampled at 300 Hz, and pre-emphasis of +6 dB/mm was applied (for all frequencies greater than 50 Hz) to ensure that higher-frequency channels were eligible for peak selection, in light of the typical −6 dB/mm attenuation of voiced speech (preemphasis has been applied in previous vocoder studies; Shannon *et al.*, 1995; Dorman *et al.*, 1997).

The noise band carriers in the ACE vocoder were filtered with either 8 or 32 dB/mm rolloff away from the center frequency, corresponding to poorer and better spectral resolution, respectively. These amounts of current spread would translate to approximately −35 and −139 dB/octave, depending on the exact reference frequency (e.g., there are approximately 0.267 octaves for every mm of cochlear space surrounding 500 Hz, 0.232 octaves per mm of space surrounding 1000 Hz, and 0.215 octaves per mm surrounding 2000 Hz) and depending on the apical or basal direction of the octave change (e.g., 3 dB/mm would result in attenuation of 12 dB per octave in the apical direction, but attenuation of 14 dB per octave in the basal direction). The filtering was done in the spatial domain rather than the octave domain to simulate the spread of excitation in a physical sense while avoiding the complications of reference frequency and direction.

The amounts of current spread used in the current vocoders are much more favorable than the values used by Bingabr *et al.* (2008), who attempted to additionally model compression of the intensity dynamic range. The filter shapes in the current study were chosen to simply have relatively more and less difficult listening conditions for the NH listeners. Otherwise, this approach was similar to the one taken by Fu and Nogaki (2004), but with the spread of excitation expressed in cochlear distance rather than octaves. The filters had constant linear attenuation in the dB/mm domain, with peaked tops. The filtering was accomplished with a custom function in Praat, as follows:

$$
\begin{aligned}
&Filter\ (formula)\ldots\\
&if\ x > 1\\
&\ldots then\ self * 10\hat{}(-(abs((\log10((x/aA)+k) * length/a)\\
&-(\log10((.cf/aA+k) * length/a)) * .rolloff.per.mm)/20)\\
&\ldots else\ self\ fi\ldots,
\end{aligned}
$$

where "length," "a," and "k" are the parameters from the classic Greenwood function. aA is "A" from the

Greenwood function, renamed because Praat does not use capital letters as first letters in variable names. ".cf" is the center frequency of the channel. "x" is the frequency across the spectrum (akin to looping through each frequency bin, for all values above 1 Hz), and "self" is the dB value at that frequency.

Figure 3 illustrates the transformation of the original speech spectrum into vocoded spectra. The 24- and 8-channel vocoders simply average the energy within each band, while the ACE vocoders pick the top eight bands and represent them with peaked synthesis channels with variable rolloff.

3. "Partial-tripolar" pulsatile vocoder

The third type of vocoder was inspired by the pulsatile vocoders used by Deeks and Carlyon (2004), Churchill *et al.* (2014), and Williges *et al.* (2015). Similar to Deeks and Carlyon (2004) and Churchill *et al.* (2014), the present vocoder capitalizes on the fact that unresolved in-phase harmonics produce periodic pulses whose rate equals the fundamental. The unfiltered carrier for the pulsatile vocoder in the current study was a harmonic complex that included components from the fundamental up to multiple closest to the Nyquist frequency of 22 050 Hz; each component was equal in intensity before filtering. The fundamental frequency of the harmonic complex was 150 Hz, which was intermediate to the fundamental frequencies of the female and male talkers in the study. Because the harmonics are all orderly and in sine phase, the *F*0 of the vocoder itself would essentially override the *F*0 of the talker whose voice is being represented. In other words, voice pitch was completely neutralized as a cue for gender with this particular vocoder. Additionally, the lower harmonics would be resolved in the typical human auditory system, meaning that they would not be pulsatile; only the upper harmonics (indexed 9 and above, or 1350 Hz and above) would carry the pulsatile envelope modulation.

The vocoder frequency-channel allocation was inspired by the Advanced Bionics device with partial tripolar stimulation mode, which is currently limited to experimental use (cf. Bierer, 2007; Landsberger *et al.*, 2012; Srinivasan *et al.*, 2013). The goal of this stimulation mode is to reduce channel interaction by focusing on the area of cochlear excitation via opposite-polarity current on electrodes flanking the stimulating electrode. The broadband harmonic complex carrier was filtered into bands centered on the default frequency bands for the tripolar configuration available in the Advanced Bionics experimental fitting software BEPS+ (see Table II, right column), with either 16 or 32 dB/mm roll off implemented by filtering in the same manner as the ACE vocoder. These values correspond to poorer and better spectral resolution, respectively. Original attempts at using 8 dB/mm (to produce stimuli parallel to the ACE vocoder) produced signals that were extremely difficult to identify, perhaps because of the sparse frequency sampling for high-frequency channels that would carry the fricatives. The 8 dB/mm vocoder was therefore abandoned in favor of an easier 16 dB/mm carrier.

Figure 4 illustrates the temporal representation of each of the vocoder styles used in this study. The pulsatile vocoders produce a more stable representation of the amplitude envelope of the vowel, although they introduce periodicity into the consonant. For the noise vocoders, there were temporal distortions consistent with the presence of inherent amplitude fluctuations within filtered noise bands (cf. Oxenham and Kreft, 2014, 2016), but no appreciable difference between the styles of processing (discrete channel/ACE) in terms of the composite temporal envelope.

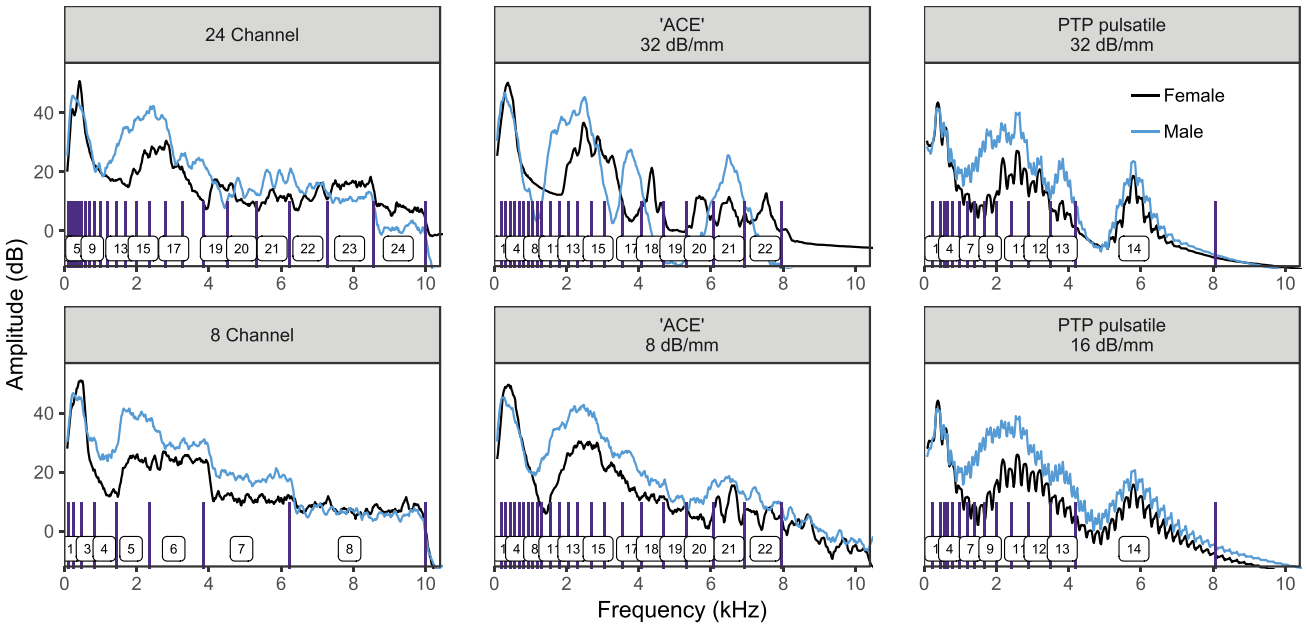

FIG. 3. (Color online) Transformation of the original spectrum of an /i/ vowel into idealized CIS-style and ACE-style vocoder signals. For the ACE vocoder, the lower two panels show spectral peak picking as well as variation in carrier filter slopes.

https://doi.org/10.1121/10.0000566

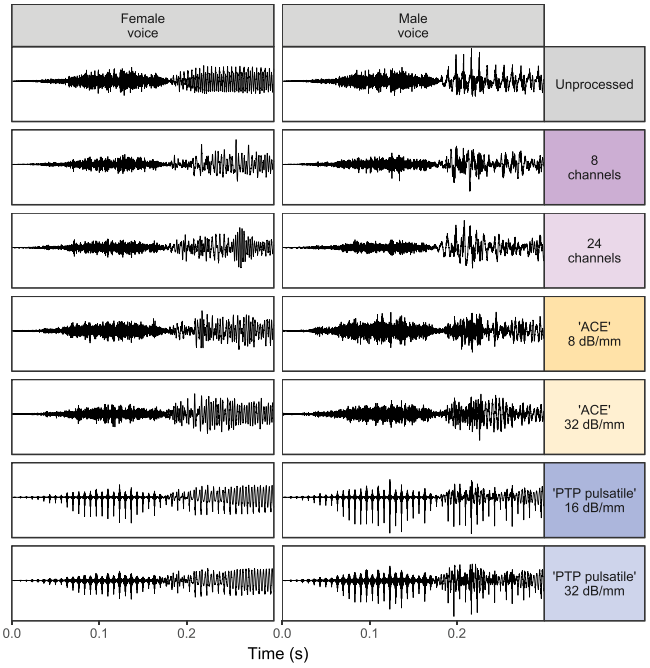

FIG. 4. (Color online) Temporal waveforms of the utterance "she" through the various vocoders used in this study. The time domain has been truncated to show detail. The top panel show unprocessed (non-vocoded) speech.

In addition to the spectral filtering and temporal envelope resolution illustrated in Figs. 3 and 4, respectively, there were some appreciable differences in the spectro-temporal contrasts that are visible in Fig. 5, which shows spectrograms from each vocoder. In particular, the formant transitions of the vowel are most easily distinguished for the ACE vocoder, where the transition of spectral energy across bands is accentuated by the dropping of channels after the transition is complete.

E. Procedure

The speech categorization task was a one-interval four-alternative forced choice procedure. After hearing a single stimulus, listeners used a computer mouse to select the word that they perceived (the choices were "see," "sue," "she," "shoe"). Although the primary interest was the perception of the binary distinction between /s/ and /ʃ/, this four-alternative paradigm was used to introduce some variety in the stimulus set, to reduce monotony and keep the task a bit more engaging.

CI listeners only heard unprocessed (natural-sounding) speech stimuli; NH listeners heard three conditions each, including unprocessed, and one of the vocoder styles presented at both levels of spectral resolution. For example, an NH listener might listen to unprocessed, 24-channel and 8-channel, or unprocessed, ACE 8 dB/mm and ACE 24 dB/mm. Data collection was done first for discrete-channel vocoders, then new participants were recruited for the ACE vocoders, and then new participants for the pulsed vocoders.

Vocoder styles were presented in blocks. For NH listeners, a single normal unprocessed speech block was always presented first. Each CI listener heard each unique stimulus four times total (or five times in the case of the seven listeners from the previous study by Winn *et al.*, 2013). Across four talkers, two vowels, and nine continuum steps, the four repetitions of each stimulus yielded a total of 288 total stimuli per condition (eight for each continuum step per talker gender), and 864 stimuli per testing session. Each NH listener heard each unique stimulus four times in each of the three test conditions. Total testing time was roughly 1.5 h (including breaks) for NH listeners and roughly 25 min for CI listeners. The presentation of tokens within each block was randomized. All testing was conducted in a sound-attenuated booth (Acoustic systems RE-143 or RE-243). A small number of listeners completed only three repetitions of some conditions because of time limitations, but psychometric functions were reliably well-formed and able to be modeled even in those cases.

Monosyllabic words were tested on a separate group of 14 young NH listeners after all of the categorization data were collected. Fifty words were played in each vocoder condition, with each block drawing words randomly sampled from a 400-word database. Listeners heard one word at a time and responded by typing the word into a computer interface.

All speech stimuli were presented at 65 dBA in the free field through a single loudspeaker. Stimulus intensity was calibrated using a Reed R8050 sound level meter using A-scale frequency weighting, at a distance of 2 feet from the loudspeaker, which is where listeners sat during testing. CI listeners who used a contralateral hearing aid removed the hearing aid during testing. No CI listener reported unaided contralateral hearing that impacted speech recognition. No attenuation devices (earplugs / headphones) were used.

Prior to testing, all listeners completed a short practice session of 12 trials to familiarize themselves with the task and interface. For NH listeners, practice blocks were performed for both unprocessed and vocoded (24-channel) speech.

F. Analysis

Listeners' responses were analyzed using a generalized linear (logistic) mixed-effects model (GLMM) in the R software interface (version 3.22, R Core Team, 2016), using the lme4 package (version 1.1–12; Bates *et al.*, 2016). The binomial family call function was used because responses were coded in a binary fashion as /ʃ/-onset (0) or /s/-onset (1). Vowel responses were coded as random effects rather than main effects, as the two vowels served only to provide variety to the stimuli. Fricative continuum step was coded in the statistical model using indices centered on step 4 (i.e., +3 for step 7, −3 for step 1; 0 for the value of 4). Step 4 was chosen as the reference because it was the one where the greatest stimulus ambiguity occurred and therefore would yield the most interpretable effects at the default

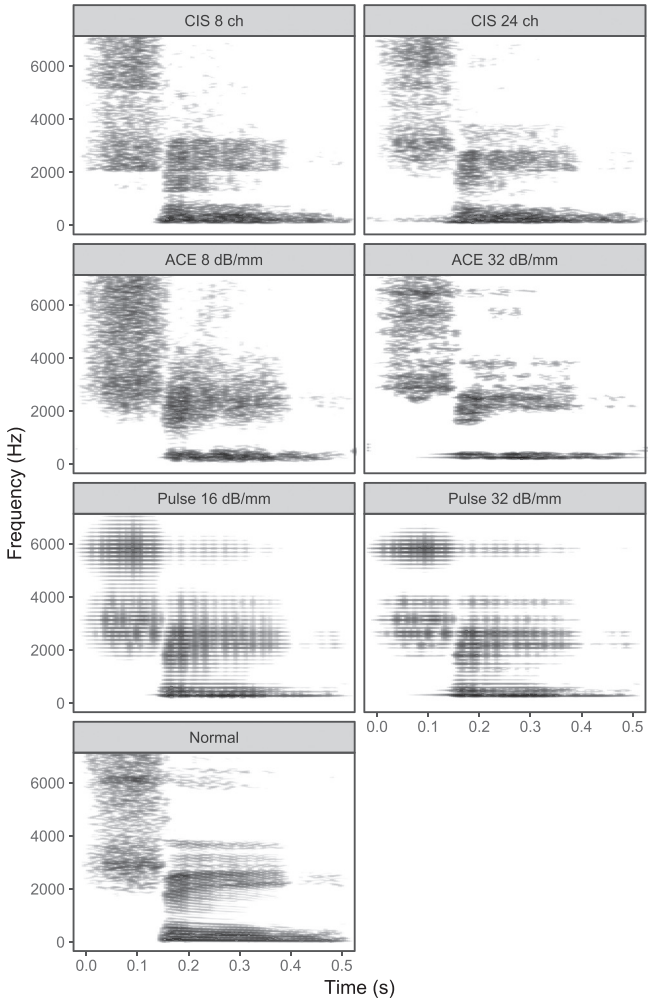

FIG. 5. Spectrograms of the word "she" processed through the vocoders used in this study.

model levels. There were two different logistic binomial models used for this study—a "complete" model used to describe all of the relevant factors in the phoneme categorization, and a "simplified" model used to specifically look at the gender-related context effect without explicitly estimating other fixed effects such as fricative step and gender-slope interactions.

The prevailing full model took the following form:

$$\begin{aligned}\text{Response} \sim\ & \text{fricative\_step} + \text{gender} + \text{condition} \\ & + \text{fricative\_step} : \text{condition} + \text{gender.condition} \\ & + \text{fricative\_step} : \text{gender.condition} \\ & + (1 + \text{fricative\_step} + \text{vowel} + \text{gender} \\ & + \text{condition}/\text{Listener})\end{aligned}$$

The model formula translates as follows: perception of /s/ was predicted by the step in the fricative continuum (slope), talker gender and condition (type of vocoder processing/unprocessed speech), the interaction between fricative step (slope) and condition, the interaction between gender and condition, and the three-way interaction between fricative step, gender and condition. The intercept term (not explicitly declared in the model) calculates the bias toward hearing /s/. Random effects (in parentheses) were intercept, fricative step (slope), vowel context, talker gender, and condition, all estimated for each listener. The supplementary material Fig. S1 illustrates the progressive increase in model accuracy with the addition of terms leading to the full model.

An *example* of context effect as an outcome measure is illustrated in Fig. 6. In the basic condition of full-resolution speech signals presented to listeners with NH thresholds, there is a clear difference in the psychometric function for phoneme categorization depending on whether fricatives are appended to a female or male voice. This difference reflects the "context effect" that will serve as the primary outcome measure for the multitude of conditions explored in this study.

III. RESULTS

A. Word recognition

Monosyllabic word recognition results are displayed in Fig. 7, revealing that all vocoders produced average results between 77% and 91%, which is in the upper range of results in modern CIs (Blamey *et al.*, 2013). For both the CIS-style and ACE-style vocoders, poorer spectral resolution resulted in poorer word recognition scores. This trend did not hold for the pulsatile vocoder, where changes in resolution did not yield any meaningful change in outcomes (89% for 32 dB/mm filters and 90% for 16 dB/mm filters). This is likely explained by the fact that the filters were both steep enough to not have a substantial impact on any phonetic cue that would be essential for recognizing a word.

An analysis of variance (ANOVA) test revealed a significant effect of condition on word recognition scores ($d.f. = 6$; $F = 26.18$; $p < 0.001$). Follow-up t-tests revealed that the effect of spectral resolution was statistically detectable for only the CIS-style vocoders (e.g., 24 channels yielded better scores than eight channels; $p = 0.004$); other comparisons of spectral resolution effects were not statistically detectable using a criterion of 0.05 and when accounting for multiple comparisons. Scores for ACE-style vocoders were better than those for CIS-style for both the 8 db/mm vs 8-channel and the 32 dB/mm vs 24-channel comparisons (both $p < 0.001$). The pulsatile vocoder yielded better scores than the ACE-style vocoder for both the 32 dB/mm comparison and the 16 dB/mm pulsatile versus 8 dB/mm noise vocoder (both $p < 0.001$). However, in general, the scores were roughly in a constrained range with limited differentiation in clinical terms.

B. Phonetic accommodation

Average response functions for each listening condition are shown in the top row of panels in Fig. 8. All functions have sigmoidal shapes with endpoints at the floor and ceiling of the response range, suggesting that endpoint categorization was reliable in all conditions.

https://doi.org/10.1121/10.0000566

Near the center of each continuum, the responses for female voices differed from those for male voices, indicating the phonetic accommodation of talker gender. This effect emerged reliably to various degrees for all vocoders except for the conventional 8-channel noise vocoder, where it was virtually absent. This lack of context effect for the 8-channel condition is consistent with the results of Winn *et al.* (2013).

Across all three types of vocoder, and particularly for the ACE-style vocoder, the phonetic accommodation effect was larger for signals with better spectral resolution, whether via more channels (in the case of 24 versus 8 channels in the conventional CIS-style channel vocoder) or narrower spread of cochlear excitation (as in the case of the "ACE" and partial tripolar simulations). This suggests that spectral resolution likely plays a role in extracting the acoustic cues necessary to drive phonetic accommodation. This would make sense because the accommodation should result from the listener's estimation of the size of the anterior resonating chamber, which is cued by resonance frequencies that demand spectral resolution.

The phonetic accommodation effect (defined as the space between response curves elicited by female and male voices) is plotted more directly in the *middle* panels of Fig. 8. Unsurprisingly, the greatest effect is observed in the center of the continuum, where the fricative itself is rather ambiguous, allowing for greater influence of secondary contextual cues. The magnitude and morphology of this direct-effect curve can be used to compare the performance of the vocoder against the data from real CI participants in the second column. A simplified representation of the direct effect (quantified by averaging the effect across the entire continuum to derive a single-point estimate) is shown on the *bottom* row of panels in Fig. 8.

C. Statistical analysis of phonetic accommodation

Table III includes the full description of the GLMM, and a validation of the model fit is illustrated in the supplemental material.[1] Among all the predictive factors included in the model, the one that is most relevant to the topic of this experiment is the effect of talker gender on the odds of perceiving /s/, and how that effect interacts with the listening condition (i.e., normal speech, vocoded speech). In other words, the crucial outcomes are the two-way interactions between gender and condition effects. For the CI group (the default group in the model), the effect of gender was statistically detectable ($p < 0.001$), with an estimated change in log odds of 1.66, which is roughly equal to a 39 percentage-point change in perception compared to the intercept (0.07 – 0.83 log odds = 32%; 0.07 + 0.83 log odds = 71%). The talker-gender context effect was considerably larger for the NH listeners (interaction effect of 1.7 for a total coefficient of 3.36; $p < 0.001$) in the unprocessed condition compared to the CI listeners. The effect of gender for NH listeners was variable across the vocoder conditions. The effects for the 32 dB/mm pulsatile vocoder, both ACE vocoders and the 24-channel vocoder were not statistically different from the effects in the CI group. The gender context effect was significantly smaller for the 8-channel vocoder compared to the corresponding effect for CI listeners ($p < 0.001$). In the 8-channel noise vocoder condition, the accommodation effect observed in the CI listeners was almost completely nullified (default term of +1.666 plus interaction term of −1.368).

The full model summary includes a number of predictive factors that were not the target outcome measure of the study, but which nonetheless are essential in thoroughly describing the data. The slope term (response to the fricative continuum steps) indicates the translation of changing a step of the fricative continuum into log odds of changing perception from /ʃ/ to /s/. Figure 8 shows that slope was shallowest for the CI listener group; slope values for the 8-channel noise vocoder and both "ACE" vocoders were not statistically different from that for the CI group. The slopes for the 24-channel noise vocoder and both pulsatile vocoders were statistically greater (i.e., steeper) than that of the CI group. The slope value for the NH listener group for normal speech was also larger than that of the CI group, and by the largest amount, compared to any other listening condition. There was an interaction between slope and talker gender; psychometric functions for labeling the fricatives were reliably steeper when appended to the female voices for numerous conditions, including for CI listeners, NH listeners (normal speech), and for the ACE vocoder with 32 dB/mm spread, and the 8-channel noise vocoder.

There was a simplified GLMM designed to test solely for effects of talker gender across conditions (with distributed random effects of the other parameters like intercept/bias, fricative step, and vowel), without explicit modeling of the other parameters. This simplified model corresponds to the bottom row of Fig. 8, where each condition produced data that collapse to a single value. In this model, the CI listeners were again the default listener group. The NH (normal speech) group showed a statistically larger context effect than the CI listeners ($p < 0.001$). The model did not identify a difference between the context effects for the CI group, the 24-channel

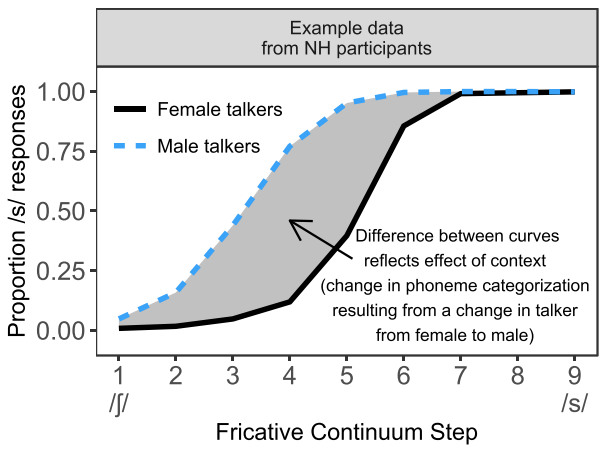

FIG. 6. (Color online) Example of "context effect" (shaded gray region) as the difference between two psychometric functions corresponding to categorization patterns for a fricative continuum that is appended to vowels produced by either female or male talkers.

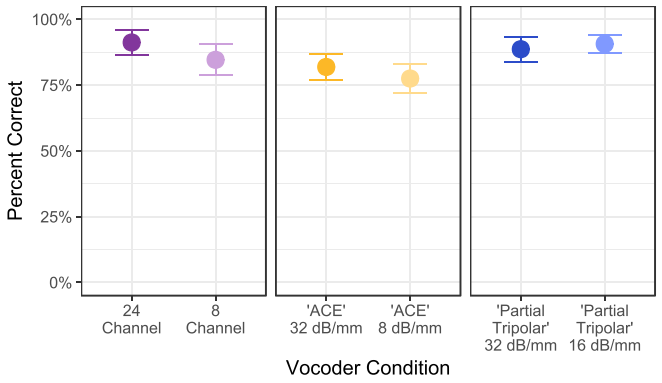

FIG. 7. (Color online) Results of monosyllabic word recognition using the six vocoder conditions in this study. Error bars indicate +/−1 standard deviation.

vocoder, and the ACE vocoder with 32 dB/mm filter slopes. Compared to the CI group, the context effect was smaller for both pulsatile vocoders (both $p < 0.001$), for the ACE vocoder with 8 dB/mm filter slope ($p < 0.05$), and dramatically smaller for the 8-channel vocoder ($p < 0.001$).

D. Individual variability in phonetic accommodation

The final method of illustrating the effect of vocoder parameters on talker accommodation is illustrated in Fig. 9, which shows the shift in phonetic boundary (in terms of units of continuum steps between an estimated 50% identification responses) across all listening conditions for all individual listeners in this study. The boundaries for each gender context were computed individually using a binomial generalized linear model (GLM) where the 50% threshold is calculated as (-Intercept/slope), which is the value of the function where log odds = 0. This method of calculating context effects is inspired by the method used by Stilp *et al.* (2015) who measured perceptual calibration to spectral filters in a speech categorization task. Rather than describing a perceptual effect in terms of log odds, this method computes the effect using continuum steps as the unit of the outcome measure. The individual data show that for the 8-channel noise vocoder, the upper end of the distribution overlaps with the lower end of the results for CI listeners. Additionally, there was one CI listener whose phonetic accommodation was remarkably higher than the rest of the group, likely due to function morphology that did not approach the floor (i.e., there was a large bias toward perceiving /s/).

IV. DISCUSSION

The results of this study demonstrate that listeners with and without cochlear implants can accommodate to talker gender when interpreting phonetic categories even when the auditory signal is degraded. As this ability is relevant to how listeners adjust to the differences in voice acoustics between women and men, we contend that it is a potentially valuable additional tool for the evaluation of vocoders and CI listeners

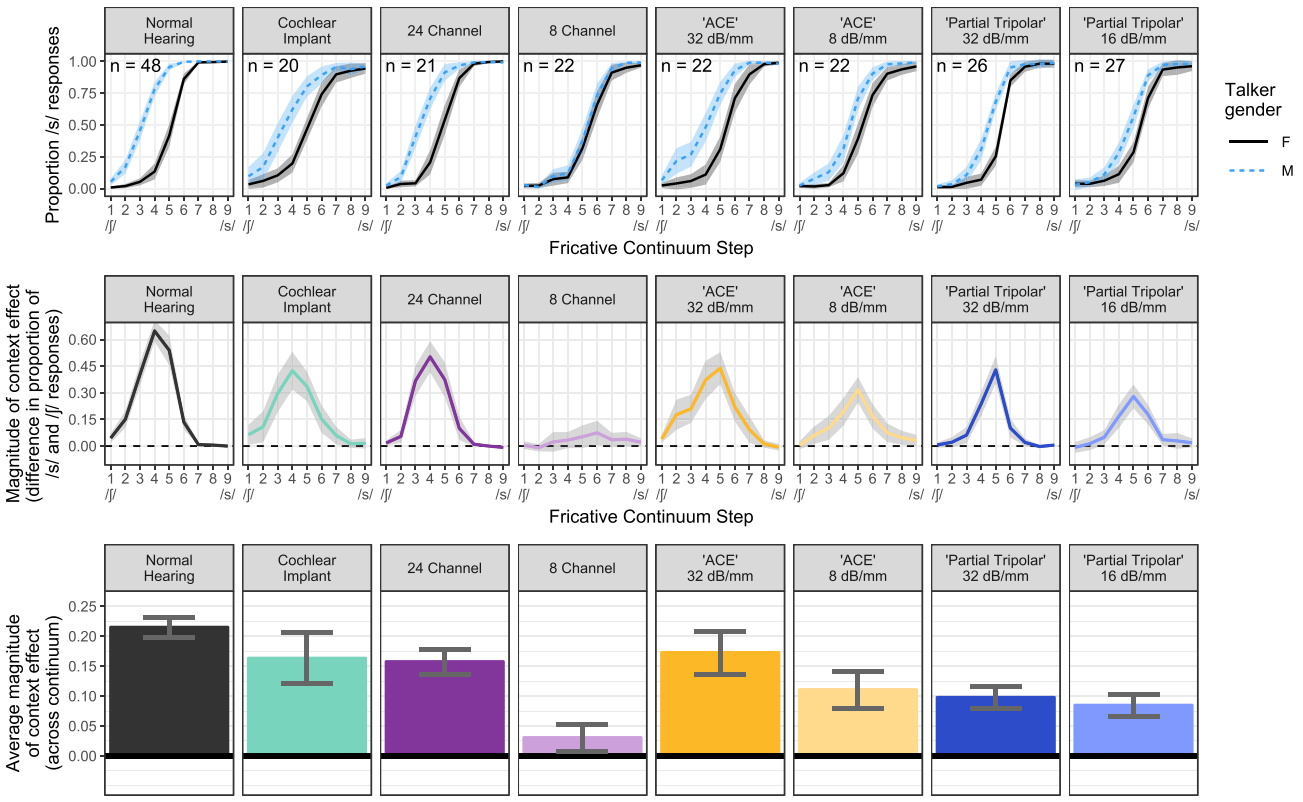

FIG. 8. (Color online) Results from all listening conditions (columns) for phoneme labeling (top row), the difference between psychometric curves (middle row), and the average magnitude of difference between curves across the entire phonetic continuum (bottom row). The middle row is a direct illustration of the effect derives from data in the top row, and the bottom row is a simplification of the data in the middle row. Error bars and width of ribbons around data indicate +/−2 standard errors of the mean.

https://doi.org/10.1121/10.0000566

in general. Different vocoders showed different amounts of success in eliciting a phonetic accommodation effect, suggesting that vocoder parameters can be adjusted to explore the auditory cues necessary to demonstrate this effect and to explore which cues might be used by CI listeners.

The use of the /ʃ/–/s/ contrast in this study was used not because it is a uniquely common or information-bearing phonetic contrast, but because it provides a well-studied example of acoustic-phonetic variation linked with talker gender. The task of phonetic accommodation of this contrast involves not only the perception of broadband aperiodic high-frequency fricative noise but also the incorporation of a low-frequency harmonic vowel segment that contains information about the talker. The task, therefore, is not suitable for the examination of one particular auditory ability but rather a complex mixture. A vocoder that best represents the information that distinguishes talker gender will likely achieve better success. The exact cues involved have not yet been fully explicated, but likely include vocal tract length and fundamental frequency (cf. Fuller *et al.*, 2014; Gaudrain and Başkent, 2018).

Toward the goal of simulating the auditory signal of a CI, the task in the current study has some advantages and some disadvantages compared to a standard test of intelligibility. The advantage is that the test is free from linguistic and cognitive influence such as lexicality and word frequency effects (since all of those factors, if present, would affect the female and male talkers equally). The test could be used in any language with a /ʃ/–/s/ contrast, and in which this phonetic distinction is not neutralized by the /i/ vowel context (as in Thai and Korean). A disadvantage is that this test has not been proven to generalize to everyday word recognition, requires precise stimulus manipulation to conduct, and is rather monotonous for the participant to perform. However, among tests that are designed to be free from

TABLE III. Summary of the generalized linear mixed-effects model. The CI listener group was the default group in the model; all model estimates are therefore to be interpreted as a deviation from the CI group estimate. Note: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; $p < 0.1$. "Fricative" corresponds to the slope of the psychometric function; "Fricative: Gender" refers to the interaction between the slope term and the gender of the talker.

Model Term	Estimate	Std. Error	*z* value	Prob (>|z|)	
Intercept (CI)	0.069	0.204	0.34	0.736	
all other effects relative to CI effect					
Intercept: Pulse 32 dB/mm	−0.762	0.236	−3.23	0.001	**
Intercept: Pulse 16 dB/mm	−0.514	0.301	−1.71	0.088	.
Intercept: "ACE" 32 dB/mm	−0.227	0.366	−0.62	0.535	
Intercept: "ACE" 8 dB/mm	−0.555	0.296	−1.87	0.061	
Intercept: 24 Channel	0.641	0.285	2.25	0.025	*
Intercept: 8 Channel	−0.473	0.301	−1.57	0.116	
Intercept: Normal Speech	0.643	0.280	2.30	0.022	*
Fricative /slope (CI)	1.166	0.132	8.82	<0.001	***
all other effects relative to CI effect					
Fricative: Pulse 32 dB/mm	0.820	0.167	4.90	<0.001	***
Fricative: Pulse 16 dB/mm	0.267	0.162	1.65	0.098	*
Fricative: "ACE" 32 dB/mm	0.302	0.165	1.83	0.067	.
Fricative: "ACE" 8 dB/mm	0.110	0.162	0.68	0.498	
Fricative: 24 Channel	0.538	0.169	3.19	0.001	***
Fricative: 8 Channel	0.106	0.162	0.65	0.514	
Fricative: Normal Speech	0.843	0.163	5.18	<0.001	***
Talker Gender (F to M) (CI)	1.666	0.244	6.83	<0.001	***
all other effects relative to CI effect					
Gender (F to M): Pulse 32 dB/mm	−0.090	0.289	−0.31	0.756	
Gender (F to M): Pulse 16 dB/mm	−0.577	0.272	−2.12	0.034	*
Gender (F to M): "ACE" 32 dB/mm	0.099	0.331	0.30	0.764	
Gender (F to M): "ACE" 8 dB/mm	−0.274	0.294	−0.93	0.352	
Gender (F to M): 24 Channel	0.363	0.333	1.09	0.275	
Gender (F to M): 8 Channel	−1.368	0.285	−4.79	<0.001	***
Gender (F to M): Normal Speech	1.698	0.330	5.14	<0.001	***
Fricative:Gender (F to M) (CI)	−0.127	0.043	−2.96	0.003	***
all other effects relative to CI effect					
Fricative:Gender (F to M): Pulse 32 dB/mm	−0.211	0.076	−2.76	0.006	**
Fricative:Gender (F to M): Pulse 16 dB/mm	0.077	0.054	1.43	0.152	
Fricative:Gender (F to M): "ACE" 32 dB/mm	−0.373	0.063	−5.92	<0.001	***
Fricative:Gender (F to M): "ACE" 8 dB/mm	0.103	0.064	1.60	0.109	
Fricative:Gender (F to M): 24 Channel	−0.204	0.084	−2.44	0.015	.
Fricative:Gender (F to M): 8 Channel	0.138	0.064	2.16	0.031	*
Fricative:Gender (F to M): Normal Speech	−0.286	0.072	−3.95	<0.001	***

linguistic influence, it has the advantage of being implemented as a test of auditory *categorization* rather than *discrimination*, rendering it immediately more relevant to speech perception in general, as argued by Holt and Lotto (2010) and Winn *et al*. (2016).

A. Spectral resolution

In three pairs of vocoders used in the current study, each had a better and poorer degree of spectral resolution. The better-resolution member of the pair always produced a larger phonetic accommodation effect. Still, the current test has not been demonstrated to be an index of spectral resolution, either in NH or CI listeners. Importantly, it is not yet clear whether this test should be able to reveal differences in performance in the range of spectral resolution thought to be representative of actual CI signals. It is also not clear whether the case of the 8-channel noise vocoder was limited in its ability to represent subtle differences in the fricative, or limited in its ability to transmit cues for talker gender within the vowel that would have been used to adjust phonetic perception. Follow-up testing with separate vocoding of the fricative and vowel could be used to disentangle these two possibilities.

In two of the three vocoder types used in the current study, spectral resolution was modified via simulated spread of excitation rather than the number of discrete channels. One advantage of changing carrier filter slope instead of number of discrete channels is that it is arguably more ecologically valid: the real-world variability in number of electrode channels is small (i.e., most of the CI participants had 22 active electrodes), but variability in factors that affect spread of neural excitation or channel interaction is large (Jones *et al*., 2013; DeVries *et al*., 2016). Litvak *et al*. (2007) showed varying degrees of speech recognition performance when holding the number of channels constant but varying simulated cochlear spread of excitation. Shannon *et al*. (1998) found that speech intelligibility in quiet was affected by the steepness of synthesis filters only when the filters were shallower than 18 dB/octave—a finding that conflicts somewhat with the results of the current study where the filters were much steeper (between −35 and −139 dB/octave when converting from the implementation of dB/mm) and still showed different effects. Vocoders with less favorable filter slopes were used by Winn *et al*. (2015) demonstrating that spectral resolution affects effort involved in listening to sentences. It is likely the case that the sensitivity of the probe stimuli is related to these differences in findings.

B. Temporal resolution and distortion

In addition to a vocoder's spectral resolution, the specific choice of carrier signals also likely plays a role in perceiving acoustic cues for gender such as *F*0 and VTL. In a noise vocoder, temporal pitch cues in the amplitude envelope should be rendered less reliable by the inherent amplitude fluctuations characteristic of filtered noise (Oxenham and Kreft, 2014, 2016). This factor might have played a role in results obtained by Fu and Nogaki (2004) and also in the ACE-style vocoders in the current study, where the resolution of filter slopes might have been inadvertently degraded by envelope fluctuations in the noise bands. Conversely, using tonal carriers minimizes these problems (cf. Grange *et al*., 2017). However, Gaudrain and Başkent (2018) surprisingly did not find improvement in *F*0 perception using vocoders with carrier envelopes explicitly designed to preserve temporal pitch cues (e.g., low-noise

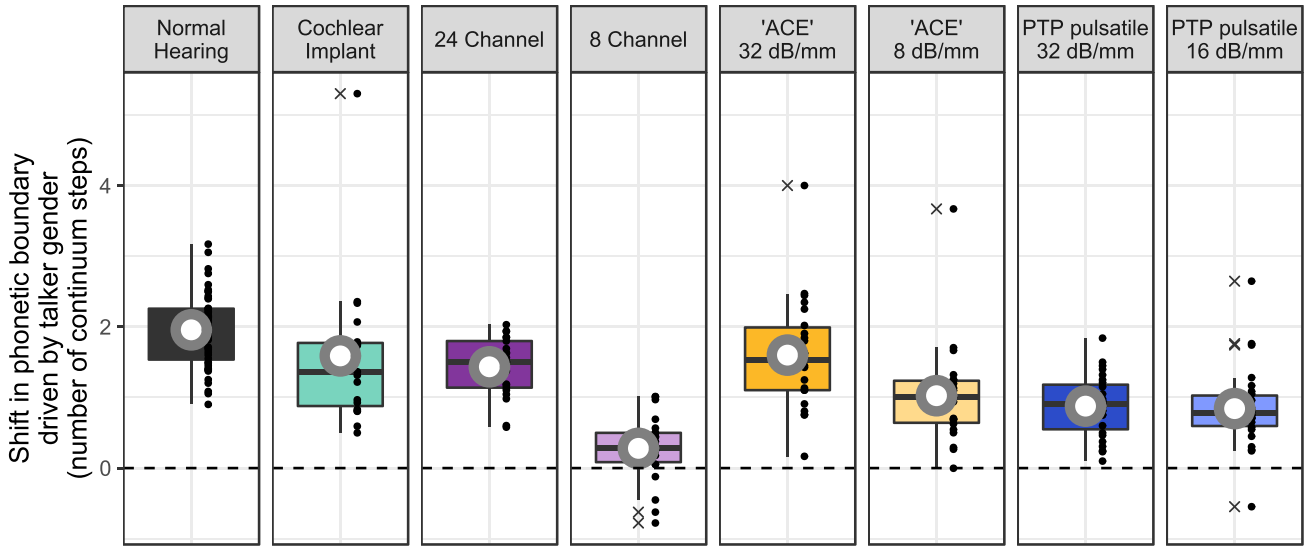

FIG. 9. (Color online) Distance (in units of fricative continuum steps) between /ʃ/–/s/ category boundaries for female and male talkers. Greater magnitude indicates a greater shift in phonetic boundary driven by talker gender. The lower and upper edges of the boxplots correspond to the first and third quartiles (the 25th and 75th percentiles), with whiskers extending from the hinge to the highest/lowest value that is within ±1.5 × range between the first and third quartiles. Data beyond the end of the whiskers are outliers plotted as Xs. All data individual data are plotted as points, and the mean within each panel is the large open circle. A single data point is omitted from a CI listener who responded with /s/ over 50% of the time for the male voice even for the lowest end of the continuum, resulting in a boundary estimate outside the range used in the experiment.

noise, pulse-spreading harmonic complexes). In their study, spectral resolution was found to play a more important role when noise carriers were used, presumably because listeners must seek cues in the form of formant frequencies of the voice as a proxy for vocal tract length since *F*0 was less reliably available.

In addition to the temporal resolution as dictated by the envelope properties of the carrier, there are complications resulting from auditory filtering in listeners with normal acoustic hearing. When frequencies of multiple carriers fall within the same auditory filter, the envelope will beat at the rate of the linear difference between those frequencies. This is the principle by which the harmonic complex vocoder in the current study neutralized the voice pitch cue—the carrier was a broadband signal where harmonics were linearly spaced at 150 Hz increments, producing a very strong percept of pitch. Grange *et al.* (2017) were able to ameliorate this problem by using a different style of carrier, where carriers were spaced according to spiral ganglion spacing rather than linear differences.

C. Perceptual learning

Of the numerous goals of the current study, the topic of perceptual learning or listening experience was not completely addressed. While it is clear that experience is not strictly necessary to demonstrate a phonetic accommodation effect with vocoded speech, those effects were obtained with signals that had spectral resolution much better than that thought to exist in real CI listeners—whether expressed as number of channels or in terms of the spread of excitation. It is therefore still a possibility that, when testing with a more realistic (poorer) resolution, that experience with the signal is necessary before demonstrating a phonetic accommodation effect. In addition to the 8-channel noise vocoder, pilot testing with a pulsatile vocoder with 8 dB/mm rolloff also yielded very little accommodation effect. The overall extremely poor phonetic labeling in the latter condition led to the abandonment of that condition, perhaps prematurely. Furthermore, other factors known to compromise CI perception such as frequency shifting, frequency compression and dynamic range compression could all demand extra listening experience as well but were not evaluated in this study.

D. Replicating CI data and previous results

One of the goals of the current study was to explore whether explicit attempts to replicate CI speech processing parameters would play a vital role in matching CI performance. The similarity of results for the 24-channel vocoder (with no such explicit attempts) and the ACE-style vocoder (with explicit processing matching) with 32 dB/mm filters suggests that the steps taken to replicate front-end processing (e.g., matching channel corner frequencies and peak-picking strategy) are not *necessary* steps to improve simulation accuracy. However, as numerous other factors are involved in the transmission of the speech signal (dynamic range compression, spectral warping and shifting, etc.), the problem is likely too complicated to have been clearly addressed by this single task. As speech can be understood despite a wide variety of distortions and simplifications (cf. Shannon *et al.*, 1995; Remez *et al.*, 1981; Saberi and Perrott, 1999), any speech recognition task might be a crude tool for examining how a signal is transmitted. The earlier study by Winn *et al.* (2013) suggested that CI listener data were better matched by NH data from unprocessed speech compared to the 8-channel vocoder data, yet one would not argue that normal hearing is a better CI simulation than the 8-channel vocoder. Clearly, the current study does not offer the final definitive answer.

The 8-channel noise vocoder results in the current study replicated the absence of phonetic accommodation observed by Winn *et al.* (2013) with a similar vocoder. There was a difference in channel-frequency allocation for the two studies. In the earlier study, the analysis filters extended up to 10 kHz, which exceeds the analysis filters of any currently available CI speech processor. The current study used a lower cutoff of 8 kHz, which provides for narrower frequency channels but also cuts off some energy in the highest-frequency spectral peak in /s/ (see Table I and Fig. 2). Despite this truncation of the /s/ spectrum, the phonetic accommodation effect in any of these vocoders is unlikely to have been influenced much by the frequency cutoff, as the frequency filtering was the same in both the female- and male-voice contexts. Additionally, similar low-pass filtering was present in the other vocoder conditions that did yield accommodation effects.

Another important aspect of CI listening that was not addressed in the current study was the upward shifting of frequency energy that is known to be very common in actual CI recipients (Holden *et al.*, 2013; Landsberger *et al.*, 2015). Even when matching frequency analysis filters of a CI speech processor, the delivery of the corresponding carrier channels is somewhat unrealistic when presented at cochlear locations that are tonotopically matched to those analysis frequencies. Upward shifting could have substantial effects in the current study because the /ʃ/ fricative would be shifted up toward the /s/ range, and the contextual information in the neighboring vocalic segment would be substantially altered as well.

E. Cue weighting

Still unknown are the perceptual mechanisms that underlie a listener's tendency to adjust to the voices of women and men. Previous work has suggested that CI and NH listeners could use different perceptual cues to categorize speech sounds (Winn *et al.*, 2012; Winn *et al.*, 2013; Winn and Litovsky, 2015; Moberly *et al.*, 2015). The results of the current study could, therefore, be interpreted to mean that perception was matched, but this conclusion is premature since listeners could have used different perceptual cues in different conditions. Recent explorations by Stilp (2017) suggest that even in spectrally degraded speech,

local spectral contrast can drive phonetic context effects. As the stimuli in this study can be characterized as local spectrally contrastive environments (e.g., the spectrum peaks in the fricative would align differently against those in the vocalic segment produced by the women and the man), it is possible that the mechanisms explored by Stilp play an important role in the phenomenon under investigation here. However, Stilp *et al.* (2015) (among others) have mainly explored spectral contrast in environments where preceding context comes *before* a target sound; in the current study, the reverse is true since the vocalic portion that serves as the context comes *after* the target fricative. As basic auditory mechanisms of contrast and enhancement might be partially explained by peripheral mechanisms that would act asymmetrically in time, it is unknown whether the same mechanisms could explain both phenomena. Other factors that could play a role are the perception of *F*0 and formant frequencies as a proxy for talker size (VTL). The current study was not designed to evaluate specific cue weighting, but ongoing work (Winn and Moore, 2019) explores these questions—cue weighting and the potential asymmetry of forward and backward context effects—using both NH and CI listeners. A tentative speculation is that *F*0 likely plays a minimal role in phonetic accommodation in the vocoder conditions in this study, since *F*0 is generally not represented well in filtered noise (despite envelope filters that maintain the fundamental frequency), and is likely overpowered by in-phase harmonic cues in the filtered harmonic complexes that comprise the pulsatile vocoder carriers.

V. CONCLUSIONS

Based on the results of this study, we conclude: (1) CI listeners can consistently accommodate to different voice acoustics of women and men when categorizing phonemes, (2) the lack of accurate simulation of the CI phonetic accommodation effect in the 8-channel noise vocoder is consistent across studies, (3) different vocoders can provide a better match to this particular aspect of phonetic perception in CI recipients, and (4) experience with degraded signals is not a strict prerequisite for showing phonetic accommodation with vocoded signals, but perceptual learning effects cannot be ruled out, since CI listeners are thought to have spectral resolution much worse than the vocoders used in this study, and might have demonstrated their phonetic accommodation as a result of experience with their implants.

In this study, stimuli crucially contained low-level aspects of speech stimuli that would not affect intelligibility, but would nonetheless influence phonetic categorization. In contrast to non-linguistic tasks such as pitch discrimination or spectral ripple discrimination, we varied acoustic cues that were inherently part of the speech signal itself, and which are relevant to the process of everyday speech perception, where listeners encounter a variety of talkers who carry differences in voice acoustics. The task of adjusting phonetic categorization according to acoustically subtle cues to talker gender is shown here to be a viable test that can be done reliably by CI users, and which can potentially distinguish vocoder simulations. The objective of the test would, therefore, be to identify a listener who presents with good intelligibility but struggles with adjusting to voices of different talkers, which is not probed in standard tests.

ACKNOWLEDGMENTS

Data collection for this study was funded by NIH-NIDCD R01 DC 004786 (M. Chatterjee), R01 DC003083 (R. Litovsky), and NIH-NIDCD: R03 DC014309 (M.B.W.). Additional support was provided by a core grant to the Waisman Center from the NIH-NICHD (Grant No. P30 HD03352). Data from the first seven CI listeners were collected with the support of the University of Maryland Center for Comparative and Evolutionary Biology of Hearing Training Grant T32 DC000046-17 (A. Popper). M.B.W. was also supported by the NIH division of loan repayment. Brianna Vandyke, Ashley Moore, Tiffany Mitchell, and Steven Gianakas assisted with data collection. Deniz Başkent, Ashley Moore, Tanvi Thakkar, and Alan Kan and three anonymous reviewers contributed helpful comments to an earlier version of this manuscript.

[1]See supplementary material at https://doi.org/10.1121/10.0000566 for visual validation of stepwise models accounting for the perception of /s/-onset words as a function of various combinations of predictors.

Aronoff, J., and Landsberger, D. (**2013**). "The development of a modified spectral ripple test," J. Acoust. Soc. Am. **134**, EL217–EL222.

Aronoff, J., Shayman, C., Prasad, A., Suneel, D., and Stelmach, J. (**2015**). "Unilateral spectral and temporal compression reduces binaural fusion for normal hearing listeners with cochlear implant simulations," Hear. Res. **320**, 24–29.

Assmann, P. (**1996**). "Modeling the perception of concurrent vowels: Role of formant transitions," J. Acoust. Soc. Am. **100**, 1141–1152.

Başkent, D., Clarke, J., Pals, C., Benard, M., Bhargava, P., Saija, J., Sarampalis, A., Wagner, A., and Gaudrain, E. (**2016**). "Cognitive compensation of speech perception with hearing impairment, cochlear implants, and aging: How and to what degree can it be achieved?," Trends Hear. **20**, 1–16.

Başkent, D., and Shannon, R. (**2005**). "Interactions between cochlear implant electrode insertion depth and frequency-place mapping," J. Acoust. Soc. Am. **117**, 1405–1416.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R., Singmann, H., Dai, B., Grothendieck, G., and Green, P. (**2016**). "Lme4: Linear mixed-effects models using 'Eigen' and S4,'" R Package Version 1.1-7, http://CRAN.R-project.org/package=lme4 (Last viewed January 14, 2020).

Bierer, J. A. (**2007**). "Threshold and channel interaction in cochlear implant users: Evaluation of the tripolar electrode configuration," J. Acoust. Soc. Am. **121**, 1642–1653.

Bingabr, M., Espinoza-Varas, B., and Loizou, P. (**2008**). "Simulating the effect of spread of excitation in cochlear implants," Hear. Res. **241**, 73–79.

Blamey, P., Artieres, F., Başkent, D., Bergeron, F., Beynon, A., Burke, E., Diller, N., Dowell, R., Fraysse, B., Gallégo, S., Govaerts, P., Green, K., Huber, A., Kleine-Punte, A., Maat, B., Marx, M., Mawman, D., Mosnier, I., O'Connor, A., O'Leary, S., Rousset, A., Schauwers, K., Skarzynski, H., Skarzynski, P., Sterkers, O., Terranti, A., Truy, E., Van de Heyning, P., Venail, F., Vincent, C., and Lazard, D. (**2013**). "Factors affecting auditory performance of postlinguistically deaf adults using cochlear implants: An update with 2251 patients," Audiol. Neurotol. **18**, 36–47.

Blumstein, S., and Stevens, K. (**1979**). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. **66**, 1001–1017.
Boersma, P., and Weenink, D. (**2011**). "Praat: Doing phonetics by computer (version 5.3.16), [computer program]," http://www.praat.org/ (Last viewed January 14, 2020).
Chatterjee, M., and Peng, S.-C. (**2008**). "Processing F0 with cochlear implants: Modulation frequency discrimination and speech intonation recognition," Hear. Res. **235**, 143–156.
Crew, J., Galvin, J., and Fu, Q.-J. (**2012**). "Channel interaction limits melodic pitch perception in simulated cochlear implants," J. Acoust. Soc. Am. **132**, 429–435.
Deeks, J. M., and Carlyon, R. (**2004**). "Simulations of cochlear implant hearing using filtered harmonic complexes: Implications for concurrent sound segregation," J. Acoust. Soc. Am. **115**, 1736–1746.
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.
DiNino, M., Wright, R., Winn, M., and Bierer, J. (**2016**). "Vowel and consonant confusion patterns resulting from spectral manipulations in vocoded stimuli designed to replicate poor electrode-neuron interfaces in cochlear implants," J. Acoust. Soc. Am. **140**, 4404–4418.
Dorman, M., Loizou, P., and Rainey, D. (**1997**). "Simulating the effect of cochlear-implant electrode insertion depth on speech understanding," J. Acoust. Soc. Am. **102**, 2993–2996.
Friesen, L., Shannon, R. V., Başkent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163.
Fu, Q.-J., Chinchilla, S., and Galvin, J. (**2004**). "The role of spectral and temporal cues in voice gender discrimination by normal-hearing listeners and cochlear implant users," J. Assoc. Res. Otolaryngol. **5**, 253–260.
Fu, Q.-J., and Nogaki, G. (**2004**). "Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing," J. Assoc. Res. Otolaryngol. **6**, 17–27.
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. Assoc. Res. Otolaryngol. **15**, 1037–1048.
Gaudrain, E., and Başkent, D. (**2018**). "Discrimination of voice pitch and vocal-tract length in cochlear implant users," Ear Hear. **39**, 226–237.
Gianakas, S., and Winn, M. (**2019**). "Perception of coarticulation in listeners with cochlear implants and other spectrally degraded conditions," J. Acoust. Soc. Am. **141**, 3839.
Grange, J., Culling, J., and Harris, N. (**2017**). "Cochlear implant simulator with independent representation of the full spiral ganglion," J. Acoust. Soc. Am. **142**, EL484–EL489.
Greenwood, D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.
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.
Holt, L., and Lotto, A. (**2010**). "Speech perception as categorization," Attn. Percept. Psychophys. **72**, 1218–1227.
Johnson, K., Strand, E., and D'Imperio, M. (**1999**). "Auditory-visual integration of talker gender in vowel perception," J. Phon. **27**, 359–384.
Jones, G., Won, J. H., Drennan, W., and Rubinstein, J. (**2013**). "Relationship between channel interaction and spectral-ripple discrimination in cochlear implant users," J. Acoust. Soc. Am. **133**, 425–433.
Jongman, A. (**1989**). "Duration of frication noise required for identification of English fricatives," J. Acoust. Soc. Am. **85**(4), 1718–1725.
Jongman, A., Wayland, R., and Wong, S. (**2000**). "Acoustic characteristics of English fricatives," J. Acoust. Soc. Am. **108**, 1252–1263.
Kovačić, D., and Balaban, E. (**2009**). "Voice gender perception by cochlear implantees," J. Acoust. Soc. Am. **126**, 762–775.
Landsberger, D., Padilla, M., and Srinivasan, A. (**2012**). "Reducing current spread using current focusing in cochlear implant users," Hear Res. **284**, 16–24.
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.
Laneau, J., Moonen, M., and Wouters, J. (**2006**). "Factors affecting the use of noise-band vocoders as acoustic models for pitch perception in cochlear implants," J. Acoust. Soc. Am. **119**, 491–506.
Litvak, L., Spahr, A., Saoji, A., and Fridman, G. (**2007**). "Relationship between perception of spectral ripple and speech recognition in cochlear implant and vocoder listeners," J. Acoust. Soc. Am. **122**, 982–991.
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 needed for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations," Psychol. Rev. **118**, 219–246.
Moberly, A., Lowenstein, J., and Nittrouer, S. (**2015**). "Word recognition variability with cochlear implants: 'Perceptual attention' versus 'auditory sensitivity,' " Ear Hear. **37**, 14–26.
Munson, B., Jefferson, S., and McDonald, E. (**2006**). "The influence of perceived sexual orientation on fricative identification," J. Acoust. Soc. Am. **119**, 2427–2437.
Oxenham, A., and Kreft, H. (**2014**). "Speech perception in tones and noise via cochlear implants reveals influence of spectral resolution on temporal processing," Trends Hear. **18**, 233121651455378.
Oxenham, A., and Kreft, H. (**2016**). "Speech masking in normal and impaired hearing: Interactions between frequency selectivity and inherent temporal fluctuations in noise," Adv. Exp. Med. Biol. **894**, 125–132.
Peterson, G., and Lehiste, I. (**1962**). "Revised CNC list for auditory tests," J. Speech Hear. Disord. **27**, 62–70.
Pals, C., Sarampalis, A., and Başkent, D. (**2013**). "Listening effort with cochlear implant simulations," J. Speech Lang. Hear Res. **56**, 1075–1084.
Patro, C., and Mendel, L. (**2016**). "Role of contextual cues on the perception of spectrally reduced interrupted speech," J. Acoust. Soc. Am. **140**, 1336–1345.
R Core Team (**2016**). "R: A language and environment for statistical computing, software version 3.3.2," R Foundation for Statistical Computing, Vienna, Austriam, https://www.R-project.org/ (Last viewed January 14, 2020).
Remez, R., Rubin, P., Pisoni, D., and Carell, T. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–950.
Saberi, K., and Perrott, D. (**1999**). "Cognitive restoration of reversed speech," Nature **398**, 760–761.
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.
Shannon, R., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Shannon, R., Zeng, F., and Wygonski, J. (**1998**). "Speech recognition with altered spectral distribution of envelope cues," J. Acoust. Soc. Am. **104**, 2467–2476.
Skuk, V., and Schweinberger, S. (**2014**). "Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender," J. Speech Lang. Hear Res. **57**, 285–296.
Srinivasan, A., Padilla, M., Shannon, R., and Landsberger, D. (**2013**). "Improving speech perception in noise with current focusing in cochlear implant users," Hear. Res. **299**, 29–36.
Stafford, R., Stafford, J., Wells, J., Loizou, P., and Keller, M. (**2014**). "Vocoder simulations of highly focused cochlear stimulation with limited dynamic range and discriminable steps," Ear Hear. **35**, 262–270.
Stilp, C. (**2017**). "Acoustic context alters vowel categorization in perception of noise-vocoded speech," J. Assoc. Res. Otolaryngol. **18**, 465–481.
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.
Williges, B., Dietz, M., Hohmann, V., and Jürgens, T. (**2015**). "Spatial release from masking in simulated cochlear implant users with and without access to low-frequency acoustic hearing," Trends Hear. **19**, 1–14.
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., 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. Psychol. **4**, 824.
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., and Moore, A. (**2019**). "Backwards and indirect context effects in accommodating gender differences in speech," in *Proceedings of the Podium Presentation at the Acoustical Society of America Spring Meeting*, May 13–17, Louisville, KY.
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**, e377–e390.
Won, J. H., Drennan, W., and Rubinstein, J. (**2007**). "Spectral-ripple resolution correlates with speech reception in noise in cochlear implant users," J. Assoc. Res. Otolaryngol. **8**, 384–392.
Xu, L., Thompson, C., and Pfingst, B. (**2005**). "Relative contributions of spectral and temporal cues for phoneme recognition," J. Acoust. Soc. Am. **117**, 3255–3267.
Zhou, N., Xu, L., and Lee, C.-Y. (**2010**). "The effects of frequency-place shift on consonant confusion in cochlear implant simulations," J. Acoust. Soc. Am. **128**, 401–409.