HHS Public Access

Author manuscript

Ear Hear. Author manuscript; available in PMC 2017 November 01.

Published in final edited form as:

Ear Hear. 2016 ; 37(6): e377–e390. doi:10.1097/AUD.0000000000000328.

Assessment of spectral and temporal resolution in cochlear implant users using psychoacoustic discrimination and speech cue categorization

Matthew B. Winn[1], **Jong Ho Won**[2], and **Il Joon Moon**[3,4]

[1]Department of Speech & Hearing Sciences, University of Washington, Seattle, WA, 98105, USA

[2]Department of Audiology and Speech Pathology, University of Tennessee Health Science Center, Knoxville, TN 37996, USA

[3]Virginia Merrill Bloedel Hearing Research Center, Department of Otolaryngology-Head and Neck Surgery, University of Washington, Seattle, WA 98195, USA

[4]Department of Otorhinolaryngology-Head and Neck Surgery, Samsung Medical Center, Sungkyunkwan University, School of Medicine, Seoul, KOREA

Abstract

Objectives—This study was conducted to measure auditory perception by cochlear implant users in the spectral and temporal domains, using tests of either categorization (using speech-based cues) or discrimination (using conventional psychoacoustic tests). We hypothesized that traditional nonlinguistic tests assessing spectral and temporal auditory resolution would correspond to speech-based measures assessing specific aspects of phonetic categorization assumed to depend on spectral and temporal auditory resolution. We further hypothesized that speech-based categorization performance would ultimately be a superior predictor of speech recognition performance, because of the fundamental nature of speech recognition as categorization.

Design—Nineteen CI listeners and 10 listeners with normal hearing (NH) participated in a suite of tasks that included spectral ripple discrimination (SRD), temporal modulation detection (TMD), and syllable categorization, which was split into a spectral-cue-based task (targeting the /ba/-/da/ contrast) and a timing-cue-based task (targeting the /b/-/p/ and /d/-/t/ contrasts). Speech sounds were manipulated in order to contain specific spectral or temporal modulations (formant transitions or voice onset time, respectively) that could be categorized. Categorization responses were quantified using logistic regression in order to assess perceptual sensitivity to acoustic phonetic cues. Word recognition testing was also conducted for CI listeners.

Results—CI users were generally less successful at utilizing both spectral and temporal cues for categorization compared to listeners with normal hearing. For the CI listener group, SRD was significantly correlated with the categorization of formant transitions; both were correlated with

Correspondence to: Il Joon Moon, MD, PhD, Department of Otorhinolaryngology-Head and Neck Surgery, Samsung Medical Center, Sungkyunkwan University, School of Medicine, Seoul, KOREA, Telephone: +82-2-3410-3579, Fax: +82-2-3410-3879, moon.iljoon@gmail.com.

Conflict of interest

The authors declare that they have no conflict of interest.

better word recognition. TMD using 100 Hz and 10 Hz modulated noise was not correlated with the CI subjects' categorization of VOT, nor with word recognition. Word recognition was correlated more closely with categorization of the controlled speech cues than with performance on the psychophysical discrimination tasks.

Conclusions—When evaluating people with cochlear implants, controlled speech-based stimuli are feasible to use in tests of auditory cue categorization, to complement traditional measures of auditory discrimination. Stimuli based on specific speech cues correspond to counterpart non-linguistic measures of discrimination, but potentially show better correspondence with speech perception more generally. The ubiquity of the spectral (formant transition) and temporal (VOT) stimulus dimensions across languages highlights the potential to use this testing approach even in cases where English is not the native language.

INTRODUCTION

Cochlear implants (CIs) are devices that can be a solution to restore hearing for people with severe to profound hearing impairment, by converting acoustic sounds into electrical signals that directly stimulate the auditory nerve. The ideal solution would be to restore perceptual acuity in both the auditory spectral (frequency) and temporal (timing/amplitude) domains. Speech (phoneme, word, or sentence) recognition in quiet is the most commonly used outcome measure for CI recipients (c.f. Lazard et al., 2012; Blamey et al., 2013). When testing perception of regular speech, however, it is not possible to isolate specific auditory abilities, because various acoustic components, such as spectral and temporal phonetic cues, as well as linguistic and contextual factors, are combined in speech, and are not neatly separable by sheer vocal control. Furthermore, it is not clear whether basic psychophysical abilities such as spectral or temporal envelope detection and discrimination in non-speech noises carry over to reflect abilities of CI users to categorize spectral or temporal specific cues that actually exist in speech, i.e. associate acoustic changes with particular speech sounds.

A classic view of auditory skills recognizes a hierarchy ranging from sound *detection* at the most basic level, moving upward to sound *discrimination*, then *identification (or categorization),* finally moving on to higher-level skills such as auditory *comprehension* (Erber, 1982). Although non-linguistic sounds (e.g. spectral ripples, pure tones, modulated noise, etc.) are capable of being discriminated, but do not fall into well-learned categories with consistent labels such as those for phonemes and words. The transfer of basic discrimination skills to higher skills like categorization is not completely understood, but it is clear that the process of speech perception is more akin to categorization than discrimination (Holt and Lotto, 2010). Toward clarifying the relationship between these two kinds of auditory abilities and their utility in predicting traditional CI speech outcome measures (such as monosyllabic word intelligibility), the present study used measures of speech cue categorization where specific spectral or temporal cues were controlled in identification tasks, while also presenting traditional psychoacoustic measures of auditory discrimination.

Spectral and temporal resolution in hearing play key roles in speech perception because speech contains a multitude of contrastive frequency and timing cues. For example, vowel contrasts and place of articulation contrasts in consonants (e.g. /b/-/d/-/g/) are signaled primarily by cues in the frequency domain, while some other contrasts like consonant voicing (e.g. /p/-/b/) and affrication (e.g. “sh” and “ch”) are associated with salient cues in the temporal domain.

In the majority of cases, people who use CIs develop functional speech recognition and report satisfaction with their devices (Holden et al. 2013). However, CIs are known to have very poor resolution in the spectral domain (i.e. the spatial spread of cochlear activity), owing to limitations in the device signal processing, the limited number of electrodes implanted in the cochlea, and the spread of neural excitation associated with electrical hearing (Boëx et al. 2003). Furthermore, CIs are limited in how deeply they can be inserted into the cochlea, resulting in upward shifting of frequency energy and further distortion to the spectral envelope, such as spectral compression (Dorman et al. 1997; Başkent & Shannon 2005).

Spectral resolution can be measured in numerous ways. For example, it is possible to ascertain the degree of spread of neural excitation resulting from the activation of each electrode (Abbas et al. 2004), and to measure the ability of CI listeners to rank pitch percepts stemming from different electrodes (Nelson et al. 1995). Another popular method of assessing spectral resolution is with an acoustic stimulus that contains a variable number of spectral peaks at a specified spectral modulation depth. Such “spectral ripple” stimuli have been used to evaluate frequency resolution in listeners with normal hearing (Supin et al. 1997), as well as listeners with hearing impairment (Henry & Turner 2003) and CIs (Henry et al. 2005). CI listeners who can discriminate ripple stimuli with a larger number of ripples per octave (i.e. a *denser* spectrum) have shown to demonstrate better speech perception in two-talker babble and steady-state noise (Won et al. 2007), and smaller degrees of electrode interaction (Jones et al. 2013).

There have been criticisms of the spectral ripple method (Azadpour & McKay 2012), particularly in terms of the lack of experimenter control over electrical stimulation, because the acoustic stimulus is transformed by the processing scheme of a CI speech processor and therefore somewhat outside the control of the experimenter. In theory, the CI processor could represent spectral peaks with varying amounts of temporal fluctuations owing to the overlaps of different spectral peaks within a single channel filter. Additionally, commonly used broadband spectral ripple stimuli cannot be used to evaluate any frequency-specific spectral resolution abilities, offering instead a coarse single-measure appraisal of resolution. Finally, spectral ripple stimuli lack the dynamic temporal structure of speech, which features fluctuations at varying rates (Rosen 1992). Despite these shortcomings, performance in spectral ripple discrimination tasks have been shown numerous times to correlate with CI listener performance on speech (Henry et al. 2005; Won et al. 2007; Anderson et al. 2011) and music (Won et al. 2011) perception, and at the very least confer a sense of how well a CI listener can discriminate stimuli that are acoustically distinct in the spectral domain. However, it should be noted that Saoji et al. (2009) found that perception of low-density spectral modulations (0.5 cycles/octave) were more strongly correlated with speech

perception than the commonly used high-density spectral ripples used in many psychophysical studies.

Under normal circumstances, it is not possible to accurately estimate spectral resolution using natural speech sounds, because speech contains a number of non-spectral cues that can affect perception. Because it is not possible to know whether speech sounds were identified specifically because of spectral cues, conventional word recognition is not a desirable test of spectral resolution in the pure sense. However, there are some spectral properties of speech that can be manipulated with sufficient accuracy such that they can be used as stimuli to probe for spectral cue perception.

For listeners with NH, among the most important spectral cues in speech are formant transitions, which are spectral peaks, usually in a harmonic complex, that correspond to particular articulatory or linguistic features. The first two (and to a lesser extent, the third) formants are known to play a crucial role in vowel perception (Hillenbrand et al. 1995). The first formant confers information about consonant voicing (Lisker 1975), and the second and third formants play a crucial role in the perception of consonant place of articulation (Dorman et al. 1977; Walley & Carrell 1983; Kewley-Port et al. 1983). It is reasonable to suspect that formants represent the most important spectral cues in speech. Accurate perception of formant cues can thus be taken as a proxy measurement of spectral resolution, at least in a *functional* sense (Winn & Litovsky 2015).

Speech contains a wide range of temporal modulations that can be broken down into general categories. For example, the syllabic rate of speech is roughly 4–7 Hz (Greenberg 1999), and the rate of modulation for individual phonemes varies widely. Consonant places of articulation can be roughly described as classes of temporal modulation envelopes (Rosen 1992), where vowels and other sonorant sounds have slow-onset envelopes, stop sounds and affricates have rapid-onset envelopes, and fricatives have envelopes somewhere in between. Consistent with this breakdown, Drullman et al. (1994) showed that stop consonant perception was particularly affected by the removal of temporal modulations above 16 Hz, while vowels were relatively more robust. In the same study, it was determined that the inclusion of temporal modulations at 16 Hz and higher added negligible benefit to overall speech reception thresholds in speech-spectrum noise, suggesting that most of the information is carried by low-frequency modulations. Elliott and Theunissen (2009) point out that the separation between formant peaks in English vowels is greater than 500 Hz (or 2 cycles/kHz, as shown in the study by Hillenbrand et al. 1995), but finer spectral resolution (up to 4 cycles/kHz) would be beneficial to detect formant transitions and other spectral details. Accordingly, Elliott and Theunissen demonstrated the dominance of temporal modulations less than 12 Hz for intelligibility in general. The similarity of consonants in terms of their respective temporal modulations across frequency channels has shown to bear relationship to the likelihood of mis-perception (Gallun & Souza 2008).

Among the finer temporal distinctions made in most languages is voice onset time, which is the time elapsed between the release of a stop consonant and the proceeding voiced segment (usually a vowel). In English, voice onset time is characterized either as short-lag (voiced) or long-lag (aspirated / voiceless), although the exact timing can vary across talkers and across

languages (Lisker & Abramson 1964). This cue has been shown to be critical in the perception of voicing in stop consonants (Abramson & Lisker 1985). Generally speaking, the boundary between voiced and voiceless stops in English is roughly 30ms (though the boundary is dependent partly on place of articulation of the consonant). The time of the modulation from burst (sound onset) to vowel onset can thus be regarded as straddling a boundary of roughly 33 Hz.

It is possible that temporal cue sensitivity is of elevated importance for CI listeners in terms of their potential success in speech perception. Compared to their NH counterparts, CI listeners are thought to rely more heavily on temporal cues for phonetic perception, on account of their poorer access to spectral cues (Winn et al. 2012). Despite the relative lesser importance of temporal modulations above 16 Hz for NH listeners, there is some evidence that CI listeners can take advantage of higher-frequency modulations. Chatterjee and Yu (2010) found a relationship between 100 Hz modulation detection and electrode discrimination, the latter of which is arguably fundamental to spectral resolution (Nelson et al. 1995). In the same group of subjects, detection of slower 10 Hz modulations at high stimulation levels was not found to be variable enough to draw conclusions about performance.

Single-electrode temporal modulation detection thresholds (MDTs) have been shown to correlate with vowel and consonant recognition in CI users (Fu 2002; Luo et al. 2008). Temporal modulation sensitivity in CI users can vary widely across different electrode stimulation sites (Pfingst et al. 2008), and the de-activation of electrodes with poor modulation sensitivity can lead to improvements in speech perception (Zhou & Pfingst 2012). In the context of electrical current spread in CIs, it is possible that modulation detection is adversely affected by spectral summation across channels. That is, single-electrode measurements of temporal modulation detection may grossly overestimate CI listeners' sensitivity to temporal cues in broadband stimulation (Galvin et al. 2014). However, Won et al. (2011) showed that broadband acoustic temporal modulation sensitivity in CI users was significantly correlated with their scores for monosyllabic word recognition and speech reception thresholds in noise, suggesting that coarse measurements can be valuable in predicting the success of CI listeners.

In this study, categorization of formant cues in speech and discrimination of spectral ripples are tested in the same population of CI patients, in order to explore the correspondence between pseudo-linguistic and psychophysical measures of spectral resolution. We say "pseudo-linguistic" because in a simple test of phonetic categorization, the goal is simple identification of the signal, rather than decoding actual linguistic content. Additionally, parallel measures were conducted to measure auditory temporal perception; voice onset time categorization was measured in conjunction with detection of temporal modulations in non-linguistic stimuli. This set of experiments was motivated by the following hypotheses: 1) perception of spectral and temporal modulations in psychophysical tasks should bear a relationship to one's ability to categorize *specific kinds* of cues in speech, rather than simply being indices of general success in speech perception. Additionally, we hypothesized that the relationship between non-linguistic psychophysical auditory discrimination would correlate with speech recognition, consistent with earlier literature (Won et al. 2007).

Finally, we hypothesized that 3) compared to non-linguistic tasks, the speech-based tasks would be stronger predictors of conventional speech recognition performance, on account of their more direct correspondence to information-bearing cues in speech.

Materials and Methods

1. Participants

Nineteen post-lingually deafened adult CI users participated in both speech cue perception and psychoacoustic tests. They were 25–83 years old (mean = 59 years, 8 males and 11 females), and all were native speakers of American English. Seven CI subjects were implanted with Advanced Bionics devices, nine subjects were implanted with Cochlear devices, and three subjects were implanted with Med-El devices. Seven of the CI participants were bilateral users; they were tested with both implants functioning. CI subjects were tested with their everyday program using their own sound processor(s) set to a comfortable listening level. CI sound processor settings were the same in both speech cue categorization and the psychoacoustic experiments. Individual CI subject information is listed in Table 1.

As a control group, ten NH subjects (18–60 years; mean age of 26 years) participated in both experiments. All NH subjects had audiometric thresholds of 20 dB HL or less at octave frequencies between 250 and 8000 Hz in both ears. NH listeners were included to give a frame of reference for "optimal" performance in these tasks, using the same exact procedure used for the CI listeners. The use of human subjects in this study was reviewed and approved by the University of Washington Institutional Review Board.

All testing sessions were conducted in a double-walled sound-attenuating (IAC) booth. Stimuli were presented to subjects via a Crown D45 amplifier and a free-standing speaker (B&W DM303) located directly in front of the subject at a distance of 1-meter. The speaker exceeded ANSI standards for speech audiometry, varying ±2 dB from 100 to 20,000 Hz. The phase response of the speaker was smooth across frequencies, varying ±30° from 150 to 20,000 Hz and ±45° below 150 Hz. Stimuli were equalized at the same root-mean-square value and presented at an average level of 65 dBA for all experiments.

2. Stimuli

2.A. Stimuli for spectral speech categorization test—The spectral speech categorization test was a replication of the procedure used by Winn and Litovsky (2015) to measure functional spectral resolution. Speech stimuli consisted of modified natural speech tokens that were spoken by a single native speaker of American English. There were three classes of sounds (corresponding to pairs of syllables) that were used for the spectral speech test, as described below. Primarily, we used /ba/ and /da/ sounds because of the prominent role of spectral cues. Additionally, fricatives /ʃ/-/s/ and approximants /r/ and /l/ were used as stimuli whose confusion with the ba-da continuum would be extremely unlikely. All speech stimuli were presented in quiet.

2.A.1 /ba/-/da/ continuum: The /ba/-/da/ continuum featured orthogonal manipulation of formant transitions and time-varying spectral tilt (i.e. the overall spectral profile, including

relative balance of high- and low-frequency energies) at the onset of the syllable. These cues (particularly formant transitions) were identified by Winn and Litovsky (2015) to be useful in measuring functional spectral resolution for CI users.

First, a continuum varying by formant transitions was created using a modification and enhancement of the basic linear predictive coding (LPC) decomposition procedure in Praat (Boersma & Weenink 2013). A schematic diagram of this formant continuum can be seen in Figure 1, which displays the lowest four formants in a simplified spectrographic representation. Only formants 2 and 3 were varied, as the first formant is ostensibly equal across these sounds; formants higher than 3 were identical across all stimuli, as they do not play a major role in the distinction of these sounds. F2 varied between roughly 1000 and 1800 Hz and transitioned to 1200 Hz 80 ms into the vowel; F3 varied between roughly 2100 and 2500 Hz and transitioned to 2300 Hz 45 ms into the vowel.

A naturally-produced /ba/ token was first down-sampled to 10,000 Hz to facilitate accurate estimation of 12 LPC coefficients. The down-sampled sound was inverse filtered to yield a residual "source" (a speech sound with no formant structure) that could be filtered using formant contours extracted from the original words. Eight contours were created, including those from the natural /ba/ and /da/ tokens, as well as 6 linearly interpolated formant contours sampled at 9 equally-spaced time points during the syllable. The consonant release bursts were filtered by the onset of the formant contour and thus complemented the formant transitions. Following the re-filtering, the signals were low-passed at 3500 Hz and added to the original signal that was high-passed above 3500 Hz (each filter had a symmetrical 1000 Hz roll-off surrounding the cutoff frequency). Thus, the full spectrum of frequency energy was preserved in the final stimuli. The high-frequency (above 3500 Hz) energy is thought to play only a negligible (if any) role in this contrast; it did not vary across the formant continuum but adds to the naturalness of the sound (Sjerps et al. 2011).

For each step in the formant continuum, spectral tilt was systematically modified in five steps in order to create orthogonal variability in a second spectral dimension. Spectra were modified using a filter that amplified or attenuated frequency energy above 800 Hz via logarithmic multiplication of the amplitude spectrum with varying slope across the frequency range. Across the spectral tilt continuum, spectra were maximally distinct at F4 (3300 Hz) and tapered to equivalent levels at 6000 Hz. At 3000 Hz, the tilt continuum ranged from +9 dB to −16 dB relative to the level of the same frequency region during the following vowel segment. The filtered stimuli were each cross-faded (over 80 ms) back into the original vowel from /ba/ so that the tilt modification affected only the consonant release/vowel onset; the uniform vowel offset neutralized any late-occurring cues to consonant identity. Each word began with a uniform short segment of pre-voicing. In sum, a total of 40 /ba/-/da/ words were created (8 formant steps × 5 spectral tilt steps). For the purpose of measuring perception of spectral cues, the formant cue alone was used as the proxy for spectral resolution (following Winn & Litovsky 2015), in accordance with its relatively uncontroversial status as the primary cue for the /ba/-/da/ contrast.

2.A.2 /ʃ/-/s/ ("sha-sa") continuum: Variability among the fricative class was introduced by creating a seven-step continuum whose endpoints were modeled after productions of /s/

and /ʃ/ sounds in natural speech. Items in the continuum contained three filtered broadband noise peaks whose center frequencies (at 2500, 5200 and 6300) aligned with proceeding formant resonances in the vowel. Consistent with the naturally-produced signals, the fricative spectrum peaks varied in terms of bandwidth, and relative amplitude. Intermediate steps in the continuum thus varied primarily by the amplitude of the fricative noise peaks relative to the formant peaks in the following vowel.

2.A.3 /ra/ and /la/ sounds: The two remaining stimuli were unaltered recordings of /ra/ and /la/ syllables which, by virtue of their consonant manner class, were predicted to be extremely unlikely to be confused for the ba/da stimuli or the fricative stimuli.

2.B Stimuli for temporal speech categorization test—Stimuli for the temporal speech categorization test varied according to voice onset time. Two twelve-step continua of voice onset time (VOT) were created – one for the p/b contrast ("pier"-"beer") and one for the t/d contrast ("tier"-"deer"). For each continuum, a natural recording of a word with voiced onset ("beer" or "deer") was used as the foundation for all stimuli. VOT took values between 4 ms and 50 ms, in roughly 4 ms increments. Following a 4 ms-long consonant onset release burst that was constant across all stimuli, 4 ms portions of the vowel onset were progressively replaced with equivalent-duration portions of the voiceless-onset aspiration sound, consistent with methods used by Andruski (1994), and McMurray et al. (2008). As such, the varying stimulus dimension was voice onset time, or the relative asynchrony of voicing relative to consonant release.

Both b-p and d-t continua were presented in the same block, and can thus be thought of as a pair of concurrent binary categorization tasks. A benefit of using two continua were that 1) stimuli were less monotonous than traditional two-choice procedures, and 2) the known differences in perceptual boundaries for /p/-/b/ (roughly 20 ms) and /t/-/d/ (roughly (35 ms) could be exploited as a secondary measure of perception of place of articulation. That is, the listeners' ability to adjust voicing perception based on consonant place of articulation could be used as an index of how well they could perceive the spectral cues that distinguished consonant place. Voiceless alveolar sounds have longer VOTs than their bilabial counterparts, and listeners adjust their categorization boundaries accordingly (Abramson & Lisker 1985). Thus, a similar trend among CI listeners would establish that they also incorporate place of articulation cues when categorizing consonant voicing.

2.C Stimuli for spectral-ripple discrimination test—Spectral ripple stimuli were generated using MATLAB with a sampling frequency of 44,100 Hz. The following equation was used:

$$s(t)=\sum_{i=1}^{2555} 10^{D\times\{abs[\sin(\pi\times R\times F_i+\varnothing)]\}/20}\times\sin\left(2\times\pi\times 50\times 100^{\frac{i-1}{2555}}\times t+\varphi_i\right), \quad \text{(Eq. 1)}$$

in which D is ripple depth in dB, R is ripples/octave, (rpo), F_i is the number of octaves at the i-th component frequency, Ø is the spectral modulation starting phase in radians, t is time in seconds, and the φ_i are the randomized temporal phases in radians for pure tones. The ripple

depth (*D*) of 13 dB was used. A total of 2555 tones were spaced equally on a logarithmic frequency scale with a bandwidth of 100 – 4991 Hz. For the reference stimulus, the spectral modulation starting phase of the full-wave rectified sinusoidal spectral envelope was set to zero radian, and for "oddball" stimulus, the phase was set to $\pi/2$ radian. The 2555 tones ensured a clear representation of spectral peaks and valleys for stimuli with higher ripple densities. The stimuli had 500 ms total duration and were ramped with 150 ms rise/fall times.

2.C1 Illustration of stimuli spectra: Figure 2 illustrates representative spectra of the speech and non-speech stimuli used in this experiment. Speech sounds (left panels) contained formant peaks that were comparable to the peaks of low-density spectral ripples (described in the next section). For speech, contrastive cues were contained in the relative frequency peak of the second resonance, whereas for ripples (right panels) the contrastive cue was peak density.

2.D Stimuli for temporal modulation detection test—The psychophysical test used to measure temporal sensitivity was modulation detection, as it is a reasonable approximation of the corresponding speech task in this study. Although some experimenters have pursued a correspondence between perception of VOT and gap detection (Wei et al. 2007; Elangovan & Stuart 2008; Mori et al. 2015), such a relationship does not logically follow from analysis of the acoustic structure of word-initial stop consonants, which do not contain a silent gap (unlike word-medial consonants and affricates). Instead, word-initial stop consonant voicing is more akin to either a modulation of the envelope or asynchrony of high- and low-frequency onsets.

Temporally modulated stimuli were generated using MATLAB software with a sampling frequency of 44100 Hz. For the modulated stimuli, sinusoidal amplitude modulation was applied to a wideband white noise carrier. The stimulus duration for both modulated and unmodulated signals were 1 second. Modulated and unmodulated signals were gated on and off with 10 ms linear ramps, and they were concatenated with no gap between the two signals.

2.D1 Illustration of stimulus temporal dimensions: Figure 3 illustrates representative time waveforms of speech and non-speech stimuli. Speech (left panels) contained a single notable modulation from low to high amplitude, essentially carrying the transition from the aspirated segment to the voiced segments in the syllable. Non-speech stimuli (right panels) contained modulations of determinate frequency that varied according to modulation depth.

2.E. Speech recognition—Monosyllabic word recognition was also conducted to facilitate comparison of the psychoacoustic and speech categorization performance to conventional measures of speech intelligibility. All CI subjects were also tested with the corpus of Consonant–Nucleus–Consonant (CNC) words used by Peterson and Lehiste (1962). Fifty CNC words were presented in quiet at 65 dBA. Word recognition was not conducted for normal-hearing listeners, as the lack of variability for such a procedure would not be informative enough to conduct correlation analysis.

3. Procedures

3.A Procedure for spectral speech categorization test—The spectral speech categorization test was performed using a 1-interval, 6-alternative forced-choice (AFC) syllable identification task. The six choices were displayed as virtual buttons on the computer screen labeled with BA, DA, SA, SHA, LA, and RA. In light of the above statements about unlikelihood of confusion across pairs, the task can be interpreted as three concurrent binary categorization tasks. The crucial contrast was between BA and DA, and the other four choices were included as "filler" trials to give variety to the stimuli, in an attempt to weaken any artificially heightened sensitivity to manipulated stimulus dimensions. Listeners were not made aware that syllables were manipulated. All listeners began with a short practice session to familiarize them with the experiment interface, the speech sounds, and the procedure. During each testing block, a total of 50 different stimuli (40 different BA-DA tokens, 8 SHA-SA tokens, RA and LA) were presented three times in random order. Each of the SHA-SA and RA-LA stimuli were repeated numerous times to roughly balance the representation of each syllable during the experiment. For each subject, three blocks of testing yielded a total of 45 observations for each step of the /bɑ/-/dɑ/ formant continuum, and a variable number of exposures to the filler trials. Total time for testing took roughly 20 minutes.

3.B Procedure for temporal speech categorization test—The temporal speech test was performed using a 1-interval, 4-AFC word categorization task. The four choices were displayed as virtual buttons on the computer screen labeled with BEER, PIER, DEER, and TIER. During each testing block, each of 24 stimuli (12 pier-beer tokens, 12 tier-deer tokens) was presented three times in random order. For each subject, three blocks of were performed; a total of 216 responses were obtained from each individual subject, with nine observations for each continuum step. All temporal speech stimuli were presented in quiet, and total testing time was roughly 15 minutes.

3.C Procedure for spectral ripple discrimination test—The spectral ripple discrimination (SRD) test was administered using a similar method as previously described by Won et al. (2007). Before actual testing, subjects listened to the stimuli several times with the experimenter to ensure that they were familiar with the task. During testing, a 3-interval, 3-AFC paradigm was used to determine the SRD threshold. Three rippled noise tokens – two reference stimuli and one "oddball" stimulus – were presented for each trial. The subject's task was to identify the interval that sounded different. Ripple density was varied adaptively in equal-ratio steps of 1.414 in an adaptive 2-up, 1-down procedure. Feedback was not provided. A level rove of 8 dB was used with a 1 dB step size to limit listeners' ability to use level cues. The threshold for a single adaptive track was estimated by averaging the ripple density for the final 8 of 13 reversals. For each subject, three adaptive tracks were completed, and the final threshold was the mean of these three adaptive tracks. Higher spectral-ripple thresholds indicate better discrimination performance. Testing time took roughly 15 minutes.

3.D Procedure for temporal modulation detection test—The TMD test was administered as previously described by Won et al. (2011). Temporal modulation detection

thresholds (MDTs) were measured using a 2-interval, 2-alternative adaptive forced-choice paradigm. One of the intervals consisted of modulated noise, and the other interval consisted of steady noise. Subjects were instructed to identify the interval that contained the modulated noise. Modulation frequencies of 10 and 100 Hz were tested: the former represents a fairly slow rate of modulation, while the other is a relatively fast rate; these rates have been used in previous investigations of temporal modulation detection by patients with atypical hearing (Shim et al., 2014; Moon et al., 2015). A 2-down, 1-up adaptive procedure was used to measure the modulation depth threshold, starting with a modulation depth of 100% and decreasing in steps of 4 dB from the first to the fourth reversal, and 2 dB for the next 10 reversals. Visual feedback with the correct answer was given after each presentation. For each tracking history, the final 10 reversals were averaged to obtain the MDT. MDTs in dB relative to 100% modulation (i.e. $20log_{10}(m_i)$) were obtained, where m_i indicates the modulation index. Subjects completed two modulation frequencies in random order, and then the subjects repeated a new set of two modulation frequencies with a newly created random order. Three tracking histories were obtained to determine the average thresholds for each modulation frequency. Testing time took roughly 15 minutes.

4. Analysis

4.A Categorization analysis—Speech cue categorization responses were analyzed toward the goal of determining the listeners' reliability and precision of translating cue levels to speech categories. Data were analyzed on the group level using generalized linear mixed-effects models (GLMMs), which are commonly used in speech categorization experiments (Morrison & Kondaurova 2009; Winn et al. 2012, 2013; Winn & Litovsky 2015; Stilp et al. 2015). The basic idea of the GLMM is to generalize methods for linear regression on a transformed response; in this case it is logistic regression using the binomial response transformed using the logit linking function. The operational outcome measure is the *slope* of the psychometric categorization function, which estimates the (log) odds that the response category (perception) will change, given a single step along the continuum parameter. The slope is expressed as a cue *coefficient*, which operates as a linear coefficient, in logit-transformed space in order to constrain outcomes (probabilities) between 0 and 1. The cue coefficient is the model estimate reported in a standard GLMM summary. Mixed effects are used to capture variability among the population sample and to explicitly model nested effects. Barr et al. (2013) provides a more detailed description of mixed-effects structure. For the formant (spectral) categorization experiment, there were random effects of listener, formant, and spectral tilt; for the VOT (temporal) categorization test, there were random effects of listener, VOT and place-of-articulation.

For the spectral speech test, response as /ba/ or /da/ was a binomial outcome measure predicted by formant continuum step and hearing, with random effects of formant slope and intercept for each subject. Perception of fricative sounds and /ra/ - /la/ sounds were not included in the analysis, since they were simply filler trials; errors on speech category (e.g. /ba/ for /sa/) were extremely rare, comprising less than 0.3% of trials). For the temporal speech cue test, response as voiceless ("pier" or "tier") or voiced ("beer" or "deer") was a binomial outcome measure predicted by VOT continuum step, place of articulation (bilabial or alveolar) and hearing status (NH or CI), with random effects of VOT slope, place of

articulation and intercept for each subject. These analyses quantify the efficiency with which the formant or VOT continua translated into perceptual changes in terms of speech categorization responses. Furthermore, the place of articulation effect in the VOT analysis provided an index of how much the listeners shifted their perceptual boundary as a product of consonant place. Higher logistic cue coefficients represented greater rate of perceptual change along the dimension being modeled, i.e. better use of formant cues in the spectral test, better use of VOT cues in the temporal test, or more adjustment to place of articulation in the temporal test.

The full data set was used to make grand statistical comparisons between the two listener groups (CI vs. NH) to assess their relative perception of the speech cues. Individual models for each listener were fit using conventional binomial logistic GLMs for the purpose of comparison against individual responses in the psychophysical tests. These individual tests took the same form as the group model, but without nested structure.

4.A Psychoacoustic data analysis—Psychoacoustic data were analyzed using conventional ANOVA testing to compare between NH and CI listener groups. Relationships between psychoacoustic results and speech categorization/ word recognition results were conducted using Pearson correlation analyses with Bonferroni correction for multiple comparisons.

RESULTS

1. Spectral Speech Categorization

Sensitivity to formant cues for NH and CI listeners was assessed by analyzing the binomial responses to the /ba/-/da/ continuum, which are summarized in Figure 4. Responses from all listeners in each group are displayed along with the standard error of the mean at each continuum step. It can be seen in this figure that NH listeners demonstrated steeper psychometric functions, consistent with an overall greater efficiency in using the formant transition cues. GLMM analysis confirmed that while CI listeners did show a significant (i.e. non-zero) use of formant cues, it was significantly weaker than that for NH listeners ($z = -10.40$; $p < 0.001$), consistent with the results of Winn and Litovsky (2015).

Two CI listeners (CI 1 and CI 14; both unilateral users) were atypical in that they did not demonstrate evidence of any contrastive categorization for /b/ and /d/ consonants. In general, their psychometric functions were flat and were extremely biased toward /d/, consistent with upward frequency shift known to occur in CI listeners. While these two listeners were among the oldest in the test group, no other demographic factors were unusual, and the older listeners (CI 6, CI 8, CI 2) did not demonstrate a similar pattern. When excluding these two participants (CI 1 and CI 14), the CI group responses showed /b/-/d/ balance that was similar to the NH group, but a relatively shallower psychometric function that was confirmed by a statistically significant reduction in the effect of formant ($z = -10.3$; $p < 0.001$).

2. Temporal Speech Categorization

Sensitivity to temporal cues in speech was assessed by analyzing psychometric functions corresponding to the different levels of VOT for the p/b continuum and the t/d continuum,

which contained differences from 4 to 50 ms duration of aperiodic aspiration (voice onset time) before a periodic vowel segment. Figure 5 illustrates two separate differences across hearing groups; responses from all listeners in each group are displayed along with the standard error of the mean at each continuum step. Compared to NH listeners, CI listeners demonstrate shallower *slopes* for both p/b and t/d continua, and they also demonstrate less *separation* between the curves corresponding to each continuum. The *slope* represents the overall efficiency in using VOT as a categorization cue, while the *separation* between curves represents the adjustment of categorization boundary attributable to consonant place of articulation. Listeners with NH showed a significantly steeper slope, observed as a larger effect of VOT, compared to CI listeners ($z = -4.17$; $p < 0.001$). NH listeners in this study replicated the often-observed trend of requiring longer VOT to categorize alveolar consonants as voiced (Lisker & Abramson 1964), observed as a significant interaction between the intercept term (i.e. overall bias) and the effect of place of articulation ($z = 8.34$; $p < 0.001$). CI listeners also demonstrated the same trend, but to a lesser degree; compared to NH listeners, the interaction between the intercept term (i.e. overall bias) and the effect of place of articulation was significantly weaker in CI listeners ($z = -4.94$, $p < 0.001$).

Two CI listeners (CI 06[bilateral] and CI 12[unilateral]) were atypical in that they did not demonstrate evidence of any contrastive voicing categorization. In general, their psychometric functions were flat and were extremely biased toward voiced sounds. They also demonstrated some of the lowest scores for word recognition and had relatively longer durations of deafness (see Table 1). When excluding these two participants, all effects remained statistically significant.

3. Spectral ripple discrimination

Performance on the SRD test was defined by the maximum number of ripples per octave (RPO) that permitted reliable discrimination of ripple phase difference. Figure 6 illustrates the results for groups and for individual listeners; NH listeners achieved a larger number of RPO (average = 5.80, s.d. = 2.12) compared to CI listeners (average = 1.40, s.d. = 1.18). A Welch two-sample t-test revealed a significant effect of hearing on the maximum RPO; $t_{13.66} = -6.32$, $p < 0.001$.

4. Temporal modulation detection

Performance for temporal modulation detection was defined by the minimum modulation depth required to reliably discriminate modulated from unmodulated noises whose modulation rates were either 100 Hz or 10 Hz. Figure 7 illustrates the results for groups and individual listeners; NH listeners could detect modulation in 10 Hz modulated noises with smaller modulation depths (average = −27.37 dB, s.d. = 1.94 dB) compared to CI listeners (average = −20.40 dB, s.d. = 5.88 dB). For 100 Hz modulated noises, NH listeners could again could detect smaller modulations (average = −19.30 dB, s.d. = 1.78 dB) compared to CI listeners (average = −7.59 dB, s.d. = 5.68 dB). Welch two-sample t-tests revealed significant differences between the NH and CI listener groups for both the 10 Hz ($t_{23.9} = 4.74$; $p < 0.001$) and the 100 Hz ($t_{23.4} = 8.30$; $p < 0.001$) conditions.

5. Relationship between speech categorization and psychoacoustic discrimination

A core motivation of this study was to establish the validity of speech categorization measures as probes of spectral and temporal resolution in CI users, supported by traditional non-linguistic measures of discrimination. Figure 8 illustrates the relationships between spectral speech and non-speech cues and temporal speech and non-speech cues. Specifically, the linguistic cues of formant transitions with spectral ripple (upper left panel) place-of-articulation (POA) with spectral ripple (lower right panel), and VOT with temporal modulation detection at 100 Hz (upper right panel) and 10 Hz (lower right panel).

SRD was significantly correlated with formant categorization ($r^2 = 0.48$; $p < 0.001$; significant following Bonferroni correction), suggesting the two tasks might tap into shared perceptual mechanisms. There was not a significant relationship between VOT perception and TMD threshold for 100 Hz-modulated noise ($r^2 = 0.178$) or 10 Hz-modulated noise ($r^2 = 0.04$), nor a significant relationship between POA and spectral ripple discrimination ($r^2 = 0.079$; $p = 0.25$) meaning the first hypothesis was only partially confirmed. The relationship between VOT and MDT at 100 Hz (Figure 8 upper right panel) approached marginal significance ($p = 0.07$), but in light of the multiple comparisons, this was not considered strong enough evidence to fully support the hypothesis.

As expected from previous studies (Won et al., 2007; 2011; Winn & Litovsky, 2015), performance of CI users was not as strong as that of NH listeners for tasks of either spectral or temporal resolution; NH listeners performed significantly better in all tasks measured in both groups in this study. Correlations between speech cue categorization and psychophysical discrimination for NH listeners was limited by the fact that performance for the categorization tasks was essentially saturated at near perfect levels (i.e. perfect separation of cue dimensions without gradiency), without much variability to covary with other abilities.

6. Relationship between cues and traditional speech recognition scores

Figure 9 displays the relationships between each of the cues hypothesized to correlate with word recognition scores (formant, VOT, POA, spectral ripple, TMD at 10 and 100Hz). The strongest relationships to speech scores were obtained for VOT (upper right panel; $r^2 = 0.51$; $p < 0.001$) and formants (upper left panel; $r^2 = 0.51$; $p < 0.001$) with SRD also being a relatively good predictor (center left panel; $r^2 = 0.44$; $p = 0.002$). The correlation between POA and word recognition (lower left panel) approached significance ($p = 0.02$), but after Bonferroni correction for multiple comparisons, this did not reach criterion for significance. TMD thresholds at 100 Hz (center right panel) and 10 Hz (lower right panel) were not found to be strong predictors of speech scores. TMD thresholds at the slower rate of 10 Hz were also not predictive of speech recognition performance in this task.

Together with the correlation tests from the previous section, all p values for these tests were Bonferroni corrected to account for 10 planned comparisons. Figure 10 illustrates all comparisons at once (ordered by strength of r-squared value) and highlights those that reached significance following correction for multiple comparisons.

DISCUSSION

Listeners with NH and with CIs were tested for categorization of spectral and temporal cues in linguistic stimuli, as well as discrimination of spectral and temporal cues in non-linguistic stimuli. We hypothesized that categorization performance in each auditory dimension would correlate to the discrimination performance in the same dimension, that psychoacoustic discrimination performance would correlate with word recognition, and that the speech cue categorization performance would be a relatively better predictor of speech recognition scores, consistent with the nature of speech perception as a process of categorization (Holt & Lotto, 2010).

Perception of both acoustic dimensions was predictably better in listeners with NH, and established a target level of performance against which CI performance can be measured. Spectral ripple discrimination had a moderately strong relationship with categorization of spectral cues in speech, partially confirming our first hypothesis. This correspondence emerged despite the lack of frequency specificity of broadband spectral ripples, and despite general lack of perfect control over acoustic-to-electric conversion in the CI speech processors. Although we hypothesized a corresponding relationship for the temporal cues, no such relationship was confirmed for cues tested here, possibly because of the difference in the inherent modulation rates (discussed later) and numbers of modulations for the speech (one) and non-speech (many) stimuli.

Spectral ripple discrimination showed correlation with speech recognition scores in CI users, replicating earlier work (Henry et al., 2005; Won et al., 2007) and supporting our second hypothesis. Stronger correspondence with word and phoneme recognition was obtained with the use of the speech categorization tests compared to the non-speech discrimination tests, consistent with the direct correspondence of the controlled speech cues with information-bearing acoustic components in the speech materials. Our third hypothesis was thus supported.

Both the linguistic and non-linguistic tests used in this study are viable tests to measure auditory abilities in the spectral and temporal domains, and each has their respective strengths and weaknesses. An advantage of using non-linguistic stimuli is that they can be used to obtain a measure of absolute acoustic sensitivity that is not affected by specific language experience. However, the /ba/-/da/ contrast is extremely common in the world's languages, including the 20 most popular languages spoken in the world, and every language with at least 50 million first-language speakers (Lewis et al. 2013).

The relationship of the psychophysical tasks to speech perception is limited by a number of factors laid out in the Introduction. For example, non-speech stimuli contain contrastive cues that might not be relevant for perception of speech, and the process of speech perception has been described as categorization rather than discrimination (Holt & Lotto, 2010). By using speech stimuli that are categorizable, the spectral and temporal speech tests in this study tapped into an auditory skill that more closely resembles everyday speech perception. In spite of the differences in spectral and temporal envelopes between the two types of stimuli,

they each tap into similar auditory abilities and are complementary methods to assess spectral and temporal resolution.

Other differences between the cues present in the linguistic and non-linguistic components of this study were the relative density and specificity of the cues in both the spectral and temporal domains. The ripple densities and temporal modulation frequencies used in the psychophysical tasks were not necessarily representative of the corresponding aspects of speech sounds. Results reported by Saoji et al (2009) underscore these observations; it was found that perception of low-density spectral modulations was more predictive of speech recognition compared to perception of high-density spectra. The spectral envelope modulation between the relevant formant peaks in the /b/ and /d/ onset spectrum are roughly on the order of 0.5 to 1.0 peaks per octave, which is well within the perceptual range of more than half of the CI listeners in this study, and close to the predictive range reported by Saoji et al. The formant peaks, however, are located in a narrow range of the frequency spectrum, as opposed to the spectral ripples, which are distributed broadly between 100 and 5000 Hz. The evenly spaced spectral peaks in the psychophysical task are not observed for most speech sounds, except for the vowel /ə/.

Although the VOT perception in the temporal speech test was not a classic temporal modulation detection test, it can be interpreted as a task of detecting a modulation corresponding to the reciprocal of the VOT boundary. That is, a VOT boundary of 30 ms is akin to a temporal modulation of 33 Hz at syllable onset. If that modulation is detected, then a voiceless perception is likely to occur. The choice of modulation depths of 10 and 100 Hz for the non-speech stimuli was motivated mainly by historical reasons, but is potentially an important limiting factor in the relevance to speech cues tested in this study. Modulation detection for 33-Hz stimuli might bear a clearer relationship to VOT perception, while performance for rates at 10 Hz and lower could correspond more closely to prosodic perception.

Another limitation of this study is the possibility that the relatively advanced age of our CI participants played an unforeseen role in the speech categorization or psychoacoustic discrimination performance. The possibility exists that the differences between NH and CI performance were attributable partly to age differences in the groups. A study by Gordon-Salant et al. (2006) confirms that older NH listeners show relatively weaker categorization responses compared to their younger NH counterparts. However, as the critical comparisons in this study did not involve cross-group comparisons (they instead were comparisons within the CI group across different tests), we are confident that the hypotheses remain unblemished by the relatively older age of the CI participant group.

Presentation of acoustic cues through a CI speech processor relinquishes some control over some of the spectral and temporal characteristics that we hope to control in studies such as the current one. Although the spectral ripple modulation densities and depths were precisely controlled in the acoustic domain, they were transformed by the processor in order to be represented using 12 to 22 electrodes. In the process, the combination of spectral filtering and compression could distort the actual spectral density and modulation depth on the “received” end of the stimulus. Similarly, the pre-emphasis that is present in most cochlear

implant processor designs is likely to over-represent higher-frequency components of broadband stimuli, rendering the TMD test either more sensitive to basal modulations than apical ones or, alternatively, relegating the basal modulations to a higher level in the dynamic range, thus subjecting it to the influence of compression. Despite this undesirable effect on acoustic-electric stimulus conversion, the status of these processor systems as the true sensory prosthetic for the participants means that the performance measured here is reflective of what performance would be in real-world acoustic situations.

It should be noted that there are a number of factors that influence speech perception other than psychophysical sensitivity to acoustic cues. For example, listeners can exploit non-acoustic knowledge about conversation context and familiarity with a talker such as dialect or speaking style. Furthermore, there are a number of factors that are likely to play a large role in the success achieved by CI listeners that is unrelated to sound coding at all. For example, the duration of deafness, age of onset of hearing impairment, degree of neural survival, implant insertion depth, and electrode-neuron interface are all factors that could affect the efficacy of a CI. Finally, conversational speech contains utterances of a much longer duration than the speech stimuli used here; normal speech perception is likely to benefit from perception of prosody and rhythm, which were not tested in this study. It may be the case that modulation detection at 10 Hz or fewer, which was not found to be predictive in this study, would indeed be predictive for longer-duration speech stimuli that contain rhythm.

In spite of the large amount of variability across CI recipients, evaluation of auditory perception in the spectral and temporal domains continues to be an important aspect of measuring the success of CIs. Performance by NH listeners in terms of perceiving and categorizing speech sounds according to various controlled cues is a potential way to set a baseline against which CI listener performance can be compared. In this study, we hypothesized that speech cue categorization would relate to performance in corresponding domains of non-speech psychoacoustic tests. Although that hypothesis was confirmed in the spectral domain, results in the temporal domain were less clear. But more importantly, we hypothesized that word recognition abilities would correspond more closely with the ability to *categorize* auditory cues (in phonetic categorization tests), instead of the ability to *discriminate* auditory cues (in non-linguistic psychoacoustic tests). This hypothesis was confirmed. Although absolute NH psychophysical sensitivity is unlikely to be matched by CI listeners with current technology, a reasonable goal could be to restore the recovery and efficient categorization of specific speech cues that are known to play a vital role in the perception of speech in general. The methods used in this study demonstrate that such testing is feasible and strongly related to word recognition abilities in CI listeners.

Acknowledgments

M.B.W. was supported by grants from the NIH-NIDCD: 1R03DC014309-01A1 (M. Winn), R01 DC003083 (R. Litovsky); R01 DC02932 (J. Edwards), a core grant to the Waisman Center from the NIH-NICHD (P30 HD03352), and by the NIH division of loan repayment. Subject compensation was supported by the educational fellowship of Advanced Bionics, donated to the laboratory of Jay Rubinstein. I.J.M. was supported by a grant of the Korean Health Technology R&D Project, Ministry of Health & Welfare (HI12C1460) and Seoul R&D Program (SS100022), Republic of Korea. Figures for this paper were produced with ggplot2 (Wickham, 2009) and Adobe Illustrator.

References

Abbas P, Hughes M, Brown C, Miller C, South H. Channel interaction in cochlear implant users evaluated using the electrically evoked compound action potential. Audiol Neurootol. 2004; 9:203–213. [PubMed: 15205548]

Abramson, A.; Lisker, L. Relative power of cues: F0 shift versus voice timing. In: Fromkin, V., editor. Phonetic Linguistics. New York, NY: Academic; 1985. p. 25-33.

Anderson E, Oxenham A, Kreft H, Nelson P, Nelson D. Comparing spectral tuning curves, spectral ripple resolution, and speech perception in cochlear implant users. J Acoust Soc Am. 2011; 130:364–375. [PubMed: 21786905]

Andruski J, Blumstein S, Burton M. The effect of subphonetic differences on lexical access. Cognition. 1994; 52:163–187. [PubMed: 7956004]

Azadpour M, McKay C. A psychophysical method for measuring spatial resolution in cochlear implants. J Assoc Res Otolaryngol. 2012; 13:145–157. [PubMed: 22002609]

Başkent D, Shannon R. Interactions between cochlear implant electrode insertion depth and frequency-place mapping. J Acoust Soc Am. 2005; 117:1405–1416. [PubMed: 15807028]

Blamey P, Artieres F, Baskent D, et al. Factors affecting auditory performance of postlinguistically deaf adults using cochlear implant: An update with 2251 patients. Audiology & Neurotology. 2013; 18:36–47. [PubMed: 23095305]

Boersma, P.; Weenink, D. Praat: doing phonetics by computer [Computer program]. 2013. Version 5.3.56, retrieved 15 Sept 2013 from http://www.fon.hum.uva.nl/praat/ (date late viewed: 8/13/2014)

Boëx C, de Balthasar C, Kós M, Pelizzone M. Electrical field interactions in different cochlear implant systems. J Acoust Soc Am. 2003; 114:2049–2057. [PubMed: 14587604]

Chatterjee M, Yu J. A relation between electrode discrimination and amplitude modulation detection by cochlear implant listeners. J Acoust Soc, Am. 2010; 127:415–426. [PubMed: 20058987]

Dorman M, Studdert-Kennedy M, Raphael L. Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. Percept Psychoph. 1977; 22:109–122.

Dorman M, Loizou P, Rainey D. Simulating the effect of cochlear-implant electrode insertion depth on speech understanding. J Acoust Soc Am. 1997; 102:2993–2996. [PubMed: 9373986]

Drullman R, Festen J, Plomp R. Effect of reducing slow temporal modulations on speech reception. J Acoust Soc Am. 1994; 95:2670–2680. [PubMed: 8207140]

Elliott T, Theunissen F. The modulation transfer function for speech intelligibility. PLoS Comput Biol. 2009; 5:e1000302: 1–14. [PubMed: 19266016]

Erber, N. Auditory Training. Washington DC: Alexander Graham Bell Association; 1982. p. 92-94.

Fu QJ. Temporal processing and speech recognition in cochlear implant users. Neuroreport. 2002; 13:1635–1639. [PubMed: 12352617]

Gallun F, Souza P. Exploring the role of the modulation spectrum in phoneme recognition. Ear Hear. 2009; 29:800–813.

Galvin J, Oba S, Fu QJ, Başkent D. Single- and multi-channel modulation detection in cochlear implant users. PLoS One. 2014; 9:e99338, 1–10. [PubMed: 24918605]

Greenberg S. Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. Speech Comm. 1999; 29:159–176.

Henry B, Turner C. The resolution of complex spectral patterns in cochlear implant and normal hearing listeners. J Acoust Soc Am. 2003; 113:2861–2873. [PubMed: 12765402]

Henry B, Turner C, Behrens A. Spectral peak resolution and speech recognition in quiet: Normal hearing, hearing impaired, and cochlear implant listeners. J Acoust Soc Am. 2005; 118:1111–1121. [PubMed: 16158665]

Hillenbrand J, Getty L, Clark M, Wheeler K. Acoustic characteristics of American English vowels. J Acoust Soc Am. 1995; 97:3099–3111. [PubMed: 7759650]

Hillenbrand J, Ingrisano D, Smith B, Flege J. Perception of the voiced-voiceless contrast in syllable-final stops. J Acoust Soc Am. 1984; 76:18–26. [PubMed: 6747105]

Holden L, Finley C, Firszt J, Brenner C, Potts, et al. Factors affecting open-set word recognition in adults with cochlear implants. Ear Hear. 2013; 34:342–360. [PubMed: 23348845]

Holt L, Lotto A. Speech perception as categorization. Attn Percept Psychophys. 2010; 72:1218–1227.

Jones G, Won JH, Drennan W, Rubinstein J. Relationship between channel interaction and spectral-ripple discrimination in cochlear implant users. J Acoust Soc Am. 2013; 133:425–433. [PubMed: 23297914]

Kewley-Port D, Pisoni D, Studdert-Kennedy M. Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. J Acoust Soc Am. 1983; 73:1779–1793. [PubMed: 6223060]

Lazard D, Vincent C, Venail F, et al. Pre-, per-, and postoperative factors affecting performance of postlingually deaf adults using cochlear implants: A new conceptual model over time. PLoS One. 2012; 7(11):e48739, 1–11. [PubMed: 23152797]

Lewis, M.; Simons, G.; Fennig, C. Ethnologue: Languages of the World. 17. SIL International; Dallas, TX: 2013. Online version: http://www.ethnologue.comLast viewed August 13, 2014

Lisker L. Is it VOT or a first-formant transition detector? J Acoust Soc Am. 1975; 57:1547–1551. [PubMed: 1141504]

Lisker L, Abramson A. A cross-language study of voicing in initial stops: Acoustical measurements. Word. 1964; 20:384–422.

Luo X, Fu QJ, Wei C, Cao K. Speech recognition and temporal amplitude modulation processing by Mandarin-speaking cochlear implant users. Ear Hear. 2008; 29:957–970. [PubMed: 18818548]

McMurray B, Aslin R, Tanenhaus M, Spivey M, Subik D. Gradient sensitivity to within-category variation in speech: Implications for categorical perception. J Exp Psych Hum Perc Perf. 2008; 34:1609–1631.

Moon IJ, Won JH, Kang HW, Kim DH, An YH, Shim HJ. Influence of tinnitus on auditory spectral and temporal resolution and speech perception in tinnitus patients. J Neurosci. 2015; 35:14260–14269. [PubMed: 26490865]

Mori S, Oyama K, Kikushi Y, Mitsudo T, Hirose N. Between-frequency and between-ear gap detections and their relation to perception of stop consonants. Ear Hear. 2015; 36:464–470. [PubMed: 25565661]

Morrison G, Kondaurova M. Analysis of categorical response data: Use logistic regression rather than endpoint-difference scores or discriminant analysis. J Acoust Soc Am. 2009; 126:2159–2162. [PubMed: 19894795]

Nelson D, Van Tasell D, Schroder A, Soli S, Levine S. Electrode ranking of 'place-pitch' and speech recognition in electrical hearing. J Acoust Soc Am. 1995; 98:1987–1999. [PubMed: 7593921]

Peterson G, Lehiste I. Revised CNC lists for auditory tests. J Speech Hear Disord. 1962; 27:62–70. [PubMed: 14485785]

Pfingst B, Burkholder-Johasz R, Xu L, Thompson C. Across-site patterns of modulation detection in listeners with cochlear implants. J Acoust Soc Am. 2008; 123:1054–1062. [PubMed: 18247907]

Rosen S. Temporal information in speech: acoustic, auditory and linguistic aspects. Phil Trans R Soc Long B Biol Sci. 1992; 336:367–373.

Saoiji A, Litvak L, Spahr A, Eddins D. Spectral modulation detection and vowel and consonant identifications in cochlear implant listeners. J Acoust Soc Am. 2009; 126:955–958. [PubMed: 19739707]

Shim HJ, Won JH, Moon IJ, Anderson E, Drennan W, McIntosh N, Weaver E, Rubinstein J. Can unaided non-linguistic measures predict cochlear implant candidacy? Otol Neurotol. 2014; 35:1345–53. [PubMed: 24901669]

Sjerps M, Mitterer H, McQueen J. Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics. Neuropsychologia. 2011; 49:3831–3846. [PubMed: 22001313]

Stilp C, Anderson P, Winn M. Predicting contrast effects following reliable spectral properties in speech perception. J Acoust Soc Am. 2015; 137:3466–3476. [PubMed: 26093434]

Supin A, Popov V, Milekhina M, Tarakanov M. Frequency-temporal resolution of hearing measured by rippled noise. Hear Res. 1997; 108:17–27. [PubMed: 9213118]

Walley A, Carrell T. Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. J Acoust Soc Am. 1983; 73:1011–1022. [PubMed: 6841809]

Wickham, H. ggplot2: Elegant graphics for data analysis. NY: Springer; 2009.

Winn M, Chatterjee M, Idsardi W. The use of acoustic cues for phonetic identification: Effects of spectral degradation and electric hearing. J Acoust Soc Am. 2012; 131:1465–1479. [PubMed: 22352517]

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

Winn M, Litovsky R. Using speech sounds to test functional spectral resolution in listeners with cochlear implants. J Acoust Soc Am. 2015; 137:1430–1442. [PubMed: 25786954]

Won JH, Drennan W, Nie K, Jameyson E, Rubinstein J. Acoustic temporal modulation detection and speech perception in cochlear implant listeners. J Acoust Soc Am. 2011; 130:376–388. [PubMed: 21786906]

Won JH, Drennan W, Rubinstein J. Spectral-ripple resolution correlates with speech reception in noise in cochlear implant users. J Assoc Res Otolaryngol. 2007; 8:384–392. [PubMed: 17587137]

Zhou N, Pfingst B. Psychophysically based site selection coupled with dichotic stimulation improves speech recognition in noise with bilateral cochlear implants. J Acoust Soc Am. 2012; 132:994–1008. [PubMed: 22894220]

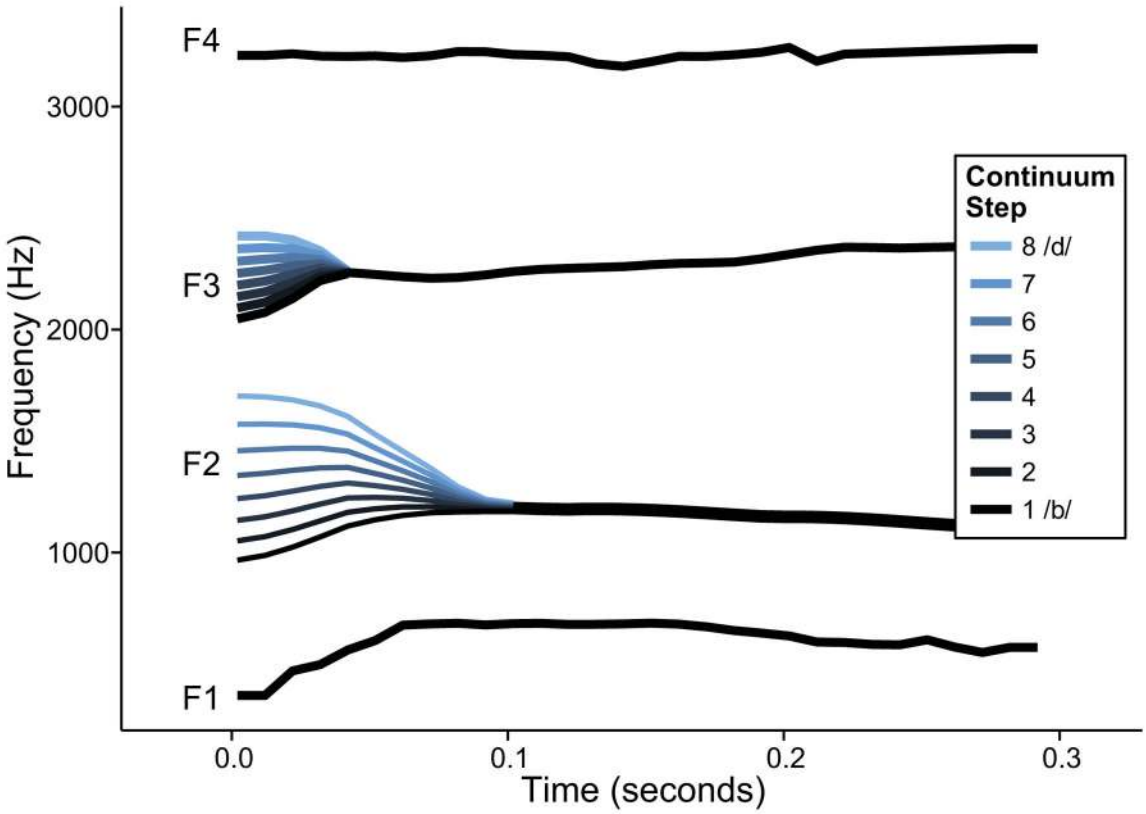

Figure 1.

Schematic spectrogram of formant contours for /ba/-/da/ stimuli in the spectral speech test. Formant transitions at word onset were varied for F2 and F3, while all other formants were equal across all steps of the continuum.

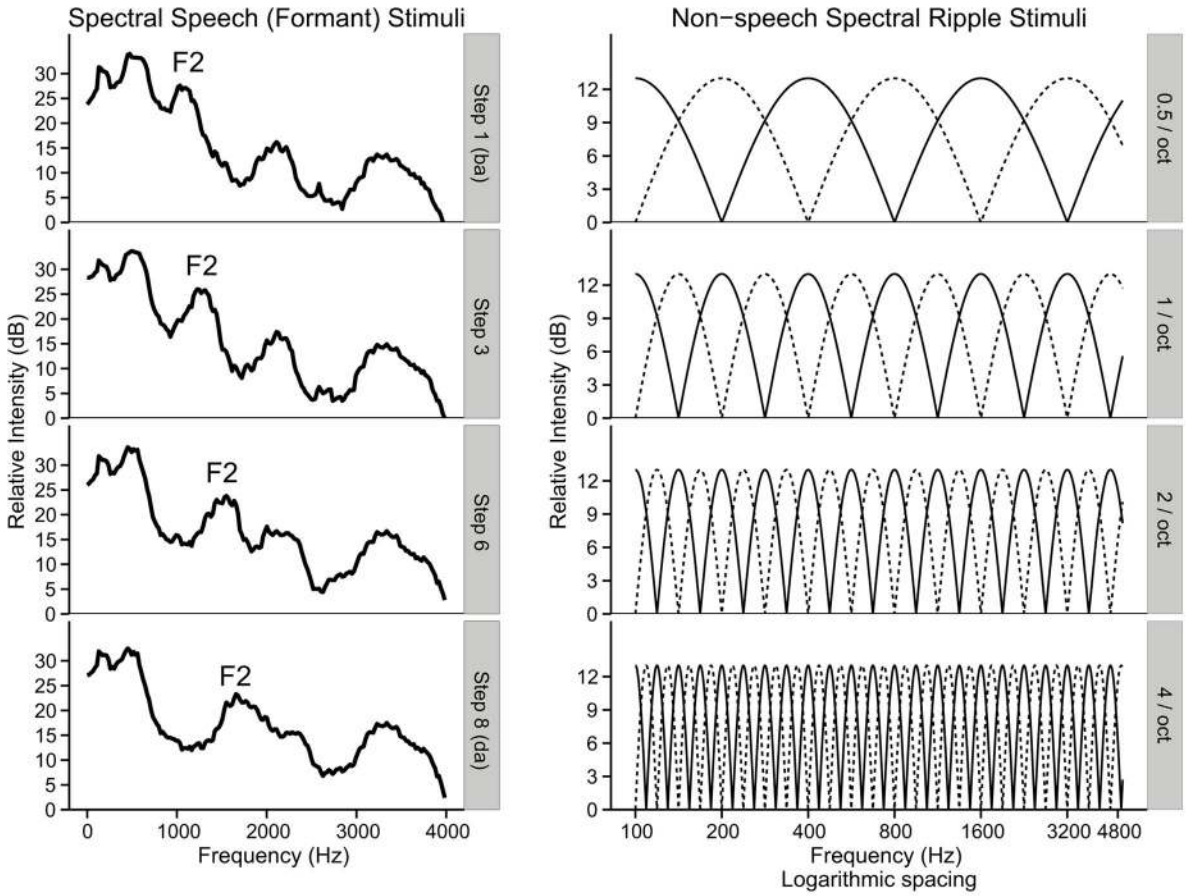

Figure 2.

Frequency-amplitude spectra for selected steps in the formant continuum (left panels) and selected spectral ripple densities (right panels). Speech spectra are obtained from the first 80 ms of each stimulus following consonant release. Dashed lines in the spectral ripple panels represent the 90-degree phase shift for each stimulus.

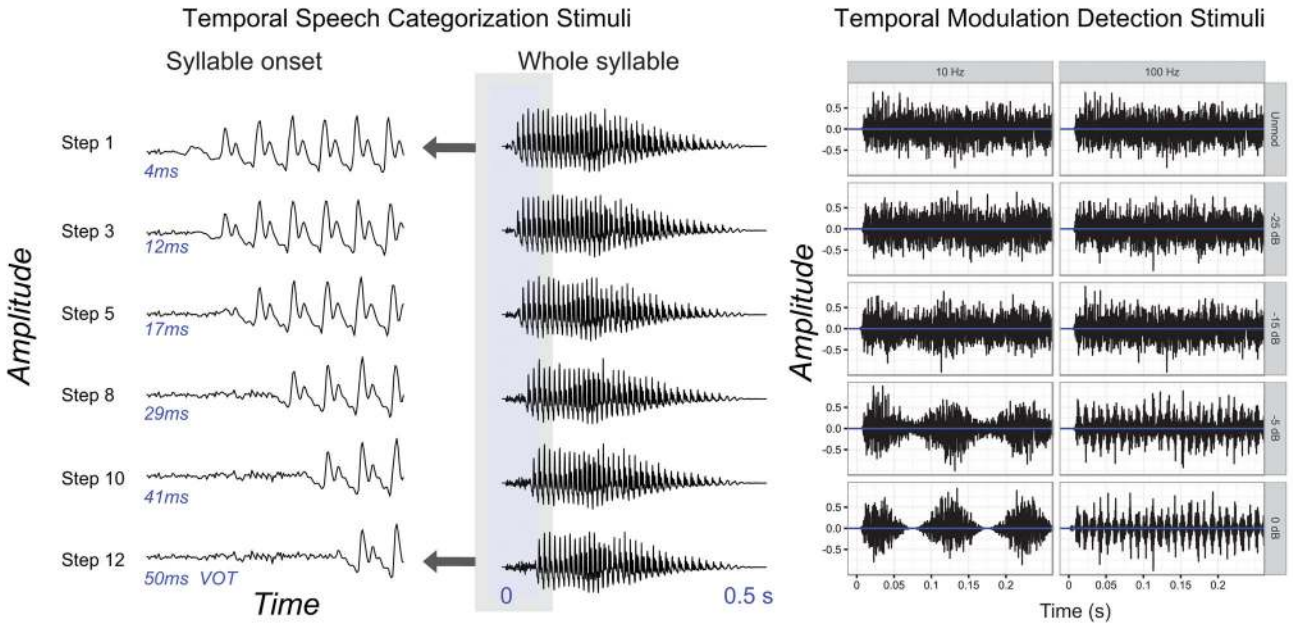

Figure 3.

Time waveforms of selected steps in the VOT continuum (left and center images) and selected modulation depths for the 10 Hz and 100 Hz modulated noise stimuli (right images). The left third of the figure illustrates a zoomed-in portion of the onset (i.e. VOT aspiration segment) of the entire syllable, which is shown in the middle portion. The right portion of the plot illustrates examples of temporal modulation detection stimuli that vary by modulation depth (rows) at two different modulation rates (columns).

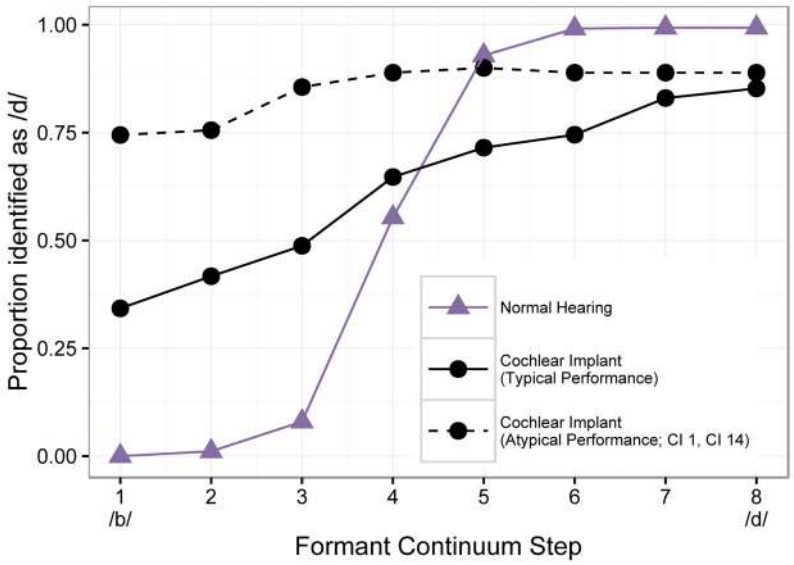

Figure 4.

Psychometric functions obtained for the formant continuum stimuli in the spectral speech categorization test. The continuum steps numbers 1 through 8 represent gradual changes from /ba/ to /da/. Dashed lines reflect data from two CI listeners (CI01 and CI14) who demonstrated especially poor perception along the continuum, labeling most stimuli as /da/. The solid line for the CI listener group excludes those two listeners.

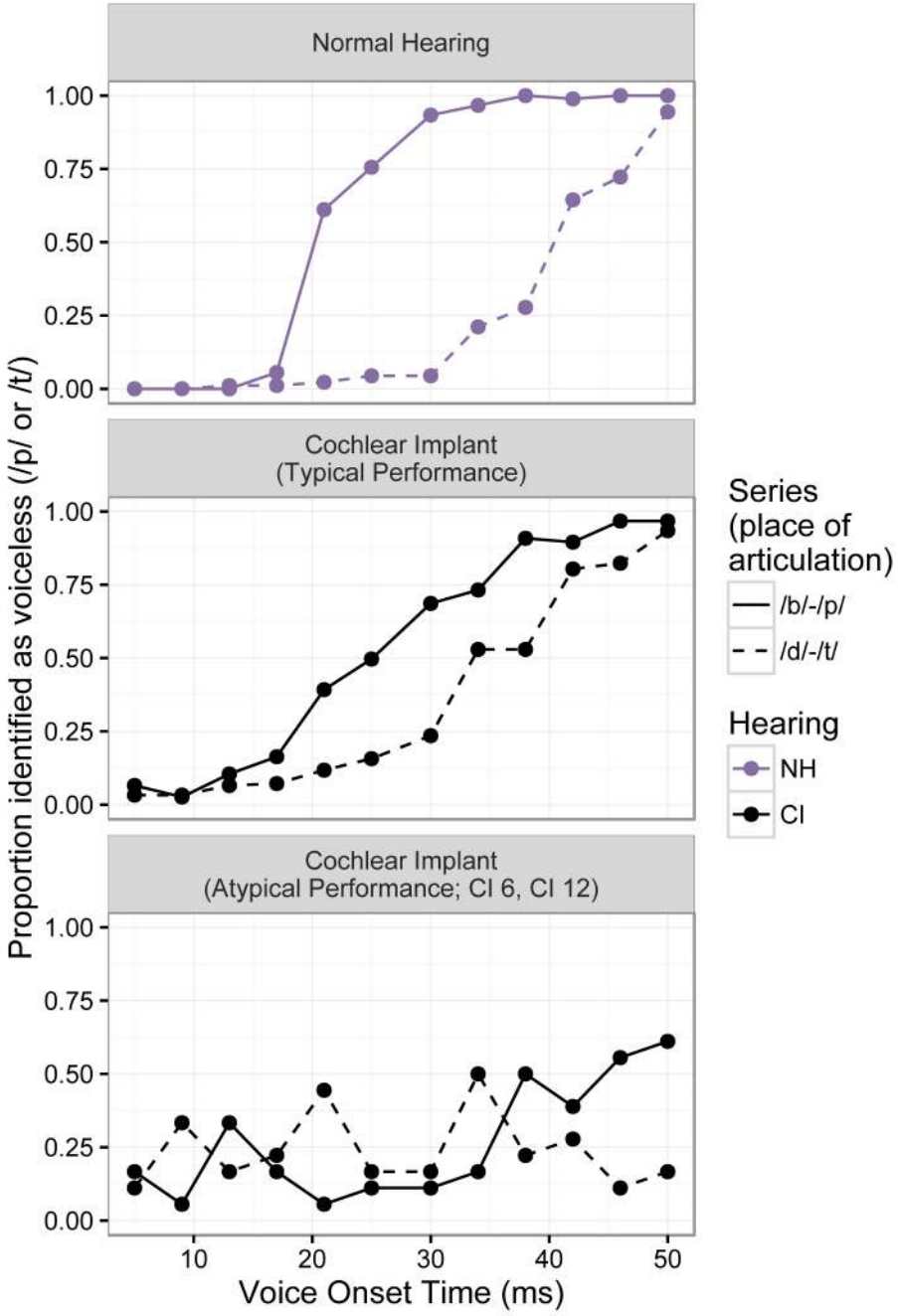

Figure 5.

Psychometric functions obtained for the VOT stimuli in the temporal speech categorization test. The continuum steps from 1 through 12 represent gradual changes from voiced (/b/, /d/) to voiceless (/p/, /t/). Solid lines correspond to perception of the p/b series, and dashed lines reflect perception of the t/d series. The space between the solid and dashed lines represents the effect of place-of-articulation (a spectral cue) on the categorization of VOT (the temporal cue).

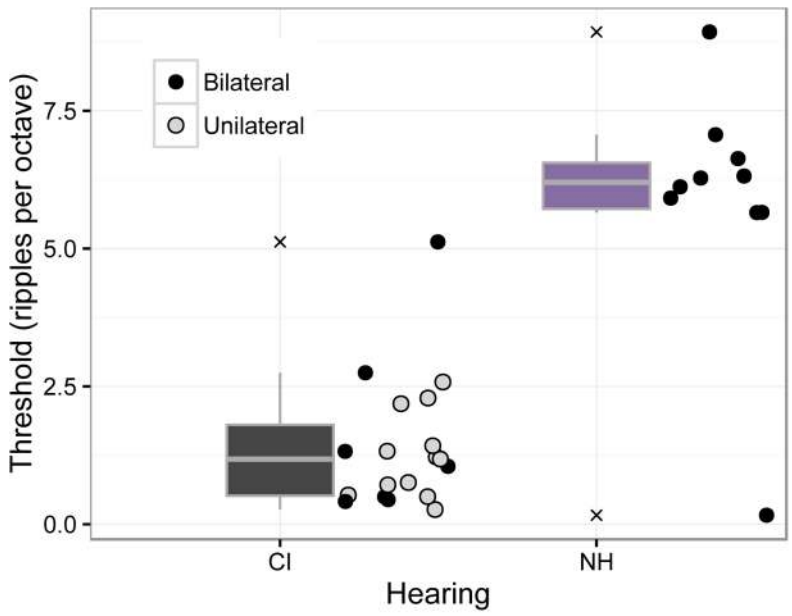

Figure 6.

Boxplots and individual data points for all listeners in the spectral ripple discrimination test. 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 * IQR of the hinge, where IQR is the inter-quartile range, or distance between the first and third quartiles. Data beyond the end of the whiskers are outliers and plotted as Xs. Within the individual points for CI users, black-filled points show data for bilateral users (tested with both processors on) and gray-filled points show data for unilateral users.

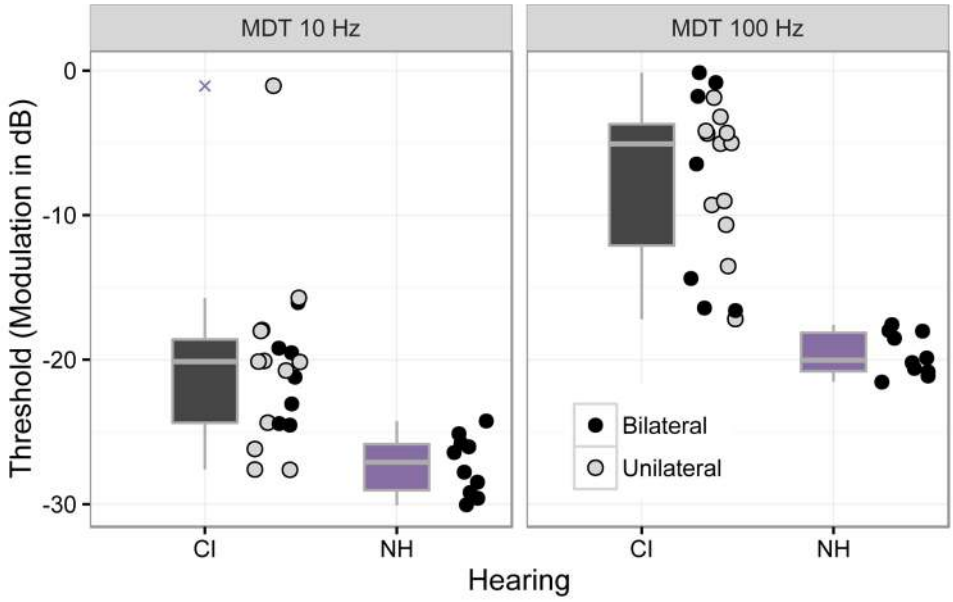

Figure 7.

Boxplots and individual data points for all listeners in the temporal modulation detection test. Boxplot dimensions are defined the same as for Figure 6.

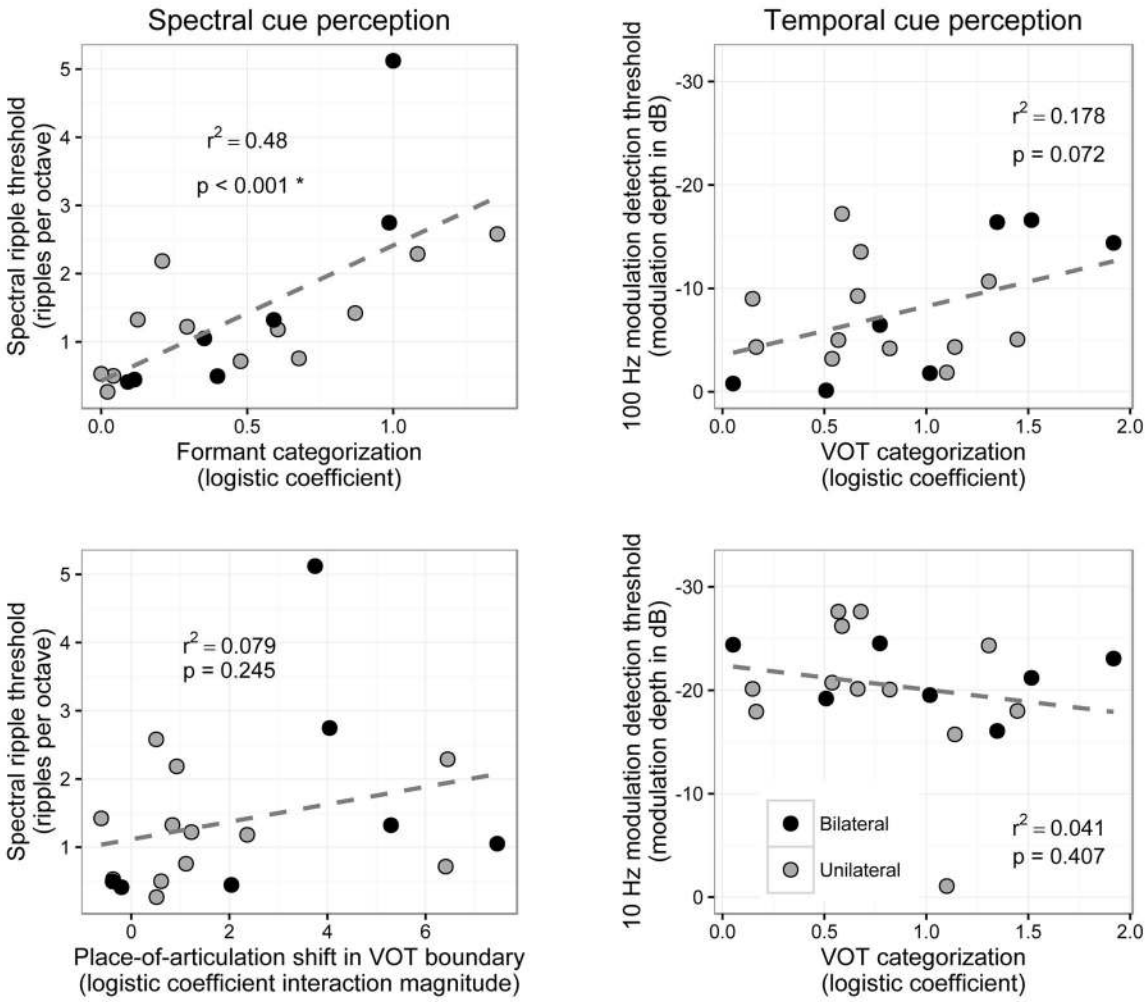

Figure 8.

Scatterplot of psychoacoustic discrimination results (y axes) as a function of speech cue categorization results (x axes), for CI listeners only. Left panels correspond to "spectral" tests and right panels correspond to "temporal" tests. The left panels show spectral ripple discrimination threshold as a function of formant cue GLM coefficient (top left panel) and place-of-articulation (POA) GLM coefficient (bottom left panel) for individual listeners. POA is a spectral cue whose impact is measured as the change in categorization function of VOT. The right panels show modulation depth threshold for 100 Hz (upper right panel) and 10 Hz (lower right panel) modulated noises as a function of VOT GLM coefficient. Dashed lines indicate linear fit to the data in each plot, collapsing both unilateral and bilateral users into a single group. The asterisk reflects significance after Bonferroni correction for ten planned comparisons (including the four on this figure and six on the next figure).

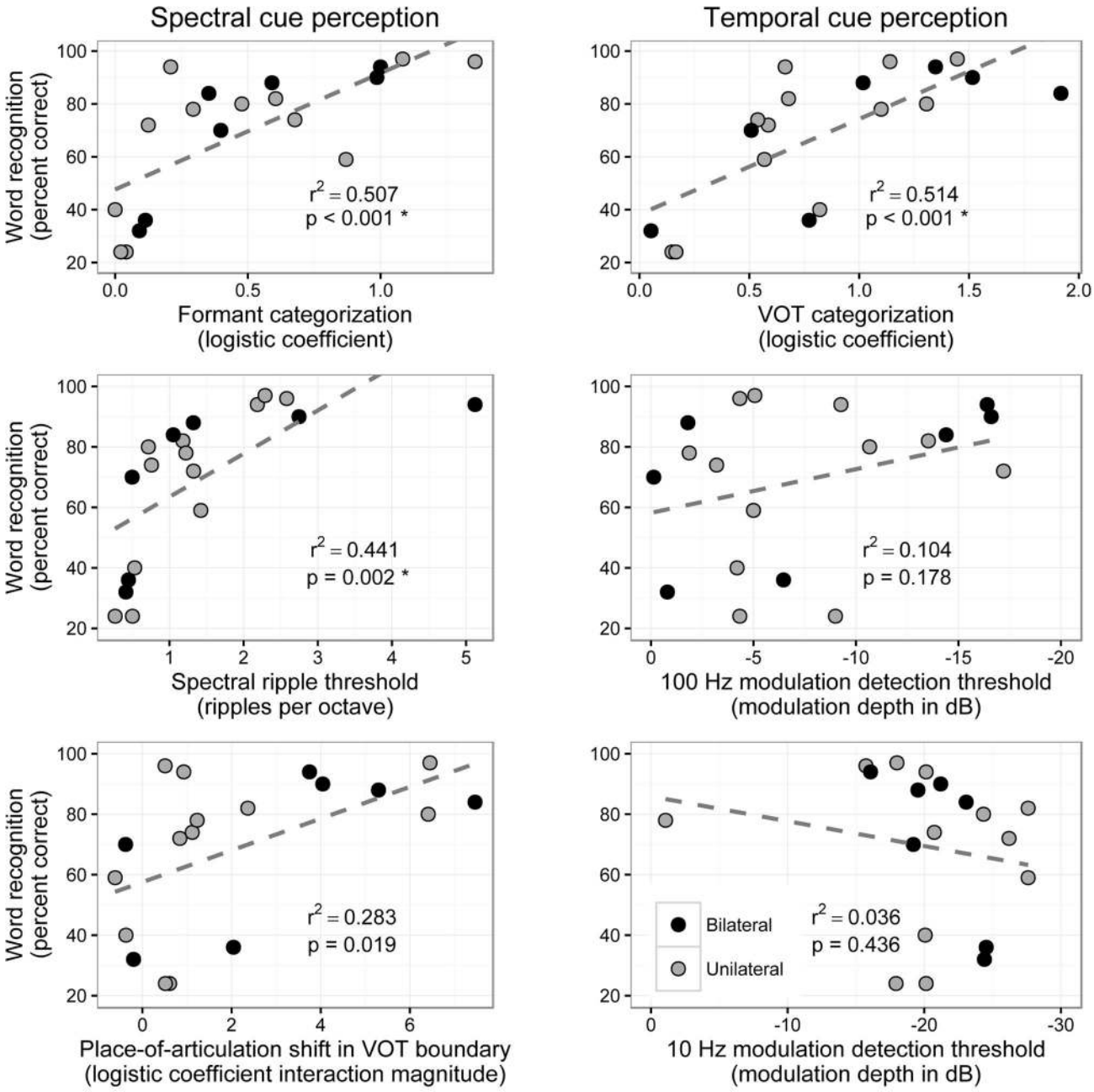

Figure 9.

Scatterplot of speech recognition scores (y axes) as a function of performance in the psychoacoustic and speech cue tests labeled on each x axis. Left panels correspond to "spectral" tests and right panels correspond to "temporal" tests. Dashed lines indicate linear fit to the data in each plot, collapsing both unilateral and bilateral users into a single group. The asterisks reflect significance after Bonferroni correction for ten planned comparisons (including the six on this figure and four on the previous figure).

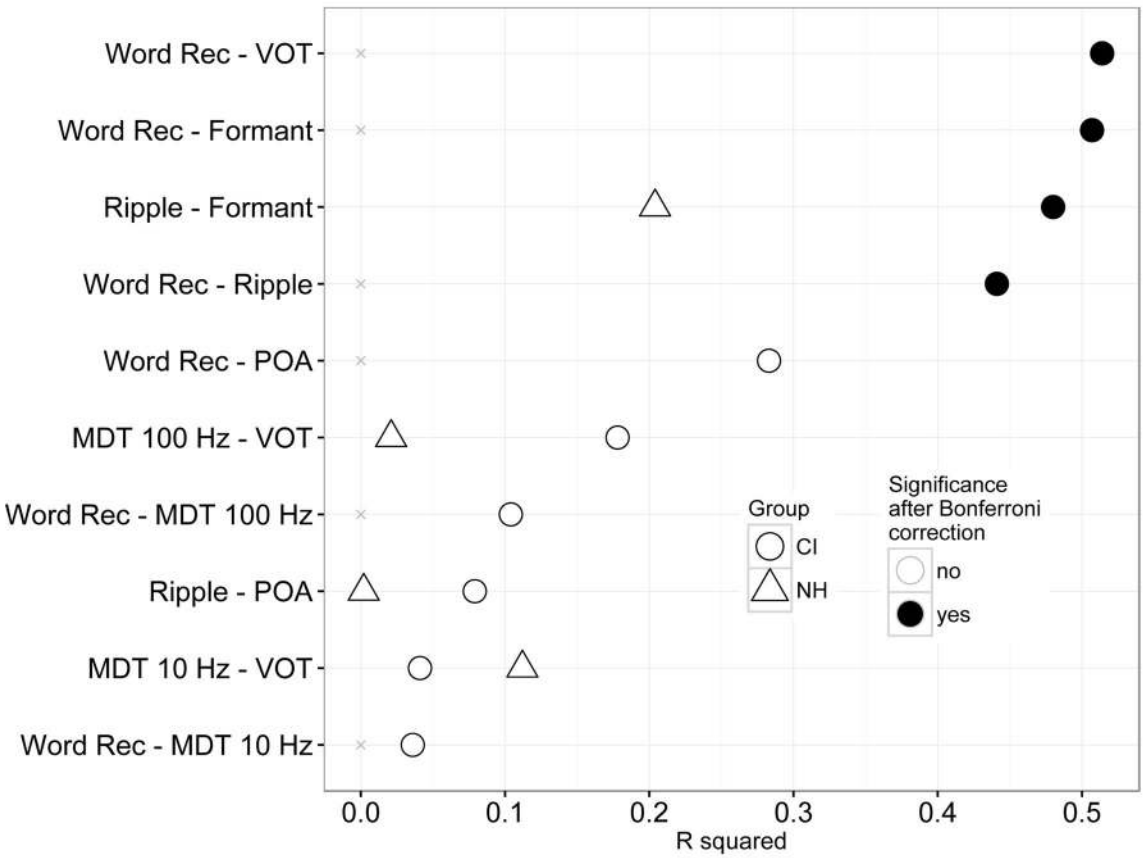

Figure 10.

R-squared correlation values for each of the comparisons across different sub-tests in this study, ordered by strength of correlation (r-squared value). Specific predictions included the correspondence between "spectral" tests (spectral ripple discrimination, formant categorization, POA adjustment), and between "temporal" tests (temporal modulation detection at 100 Hz and 10 Hz, and VOT categorization. For each comparison, the p-value was adjusted using the Bonferroni correction for ten planned comparisons; statistical significance after this correction is indicated by solid black fill in the data points. Listeners with NH did not complete word/phoneme recognition, and therefore have no correlations to report for any comparisons involving those two metrics.

Table 1

Cochlear implant subject demographics. Duration of CI use is included in duration of hearing loss.

Subject ID	Age at testing (yrs)	[a] Duration of hearing loss (yrs)	Duration of CI use (yrs)	Etiology	Implant type	CNC word	CNC phoneme	Speech processing strategy
CI 1	73	25	7	Unknown	HiRes90K	72	88	HiRes
CI 2	83	35	6	Unknown	HiRes90K	59	77	Fid.120
CI 3	25	20	10	[b] EVA syndrome	Freedom (Bi)	90	94	ACE
CI 4	70	20	9	Genetic	Freedom	74	89	ACE
CI 5	57	57	9	Connexin 26	Combi40+ (Bi)	36	60	CIS
CI 6	74	40	13	Genetic	HiRes90K (Bi)	32	32	Fid.120
CI 7	49	30	3	Genetic	HiRes90K	24	45	Fid.120
CI 8	65	35	3	Unknown	Nucleus 5	96	99	ACE
CI 9	68	7	5	Genetic	Freedom	82	89	ACE
CI10	50	50	20	Hereditary	HiRes90K (Bi)	84	92	Fid.120
CI11	43	29	2	Genetic	Freedom (Bi)	94	98	ACE
CI12	50	46	1	Genetic	Freedom	24	39	ACE
CI13	40	40	0.8	Unknown	Flex28	40	63	FSP
CI14	69	35	7	Unknown	HiRes90K	94	98	Fid.120
CI15	50	17	5	Unknown	HiRes90K (Bi)	70	83	Fid.120
CI16	53	20	2	Unknown	Nucleus 5	97	99	ACE
CI17	67	50	2	Unknown	Nucleus 24 (Bi)	88	96	ACE
CI18	78	20	5	Genetic	Freedom	78	87	ACE
CI19	68	3	0.6	Genetic	Flex28	80	89	FSP

[a] Duration of severe to profound hearing loss is based on subjects' self-report of the number of years they were unable to understand people on the telephone prior to implantation.

[b] EVA: Enlarged vestibular aqueduct.

[c] Fid.120 = Fidelity 120; ACE = Advanced Combination Encoder; FSP = Fine Structure Processing