Benefits of preserving stationary and time-varying formant structure in alternative representations of speech: Implications for cochlear implants

Susan Nittrouer,[a] Joanna H. Lowenstein, Taylor Wucinich, and Eric Tarr
Department of Otolaryngology, The Ohio State University, 915 Olentangy River Road, Suite 4000, Columbus, Ohio 43212

(Received 19 February 2014; revised 18 July 2014; accepted 28 August 2014)

Cochlear implants have improved speech recognition for deaf individuals, but further modifications are required before performance will match that of normal-hearing listeners. In this study, the hypotheses were tested that (1) implant processing would benefit from efforts to preserve the structure of the low-frequency formants and (2) time-varying aspects of that structure would be especially beneficial. Using noise-vocoded and sine-wave stimuli with normal-hearing listeners, two experiments examined placing boundaries between static spectral channels to optimize representation of the first two formants and preserving time-varying formant structure. Another hypothesis tested in this study was that children might benefit more than adults from strategies that preserve formant structure, especially time-varying structure. Sixty listeners provided data to each experiment: 20 adults and 20 children at each of 5 and 7 years old. Materials were consonant-vowel-consonant words, four-word syntactically correct, meaningless sentences, and five-word syntactically correct, meaningful sentences. Results showed that listeners of all ages benefited from having channel boundaries placed to optimize information about the first two formants, and benefited even more from having time-varying structure. Children showed greater gains than adults only for time-varying formant structure. Results suggest that efforts would be well spent trying to design processing strategies that preserve formant structure.
[http://dx.doi.org/10.1121/1.4895698]

PACS number(s): 43.66.Ts, 43.71.Ft, 43.71.Es [DB] Pages: 1845–1856

I. INTRODUCTION

Advances in signal processing made toward the end of the twentieth century expanded the available options for presenting signals through sensory aids to patients with hearing loss (Levitt, 1991). In particular, high-speed digital processing made it possible to extract specific phonetic or articulatory features from the speech signal in real time, and represent them in either hearing aids (e.g., Falkner *et al.*, 1992; Fourcin, 1990) or cochlear implants (e.g., Blamey *et al.*, 1985; Tye-Murray *et al.*, 1990; Wilson *et al.*, 1991). However, as promising as the idea was, results were not as good as would have been desired with those early devices, especially cochlear implants; speech recognition equivalent to that of listeners with normal hearing was not achieved. There were likely many reasons for why that was, including problems in extracting speech features (or structure) veridically from the signal, and in presenting that structure within the constraints of the technology available at that time. Consequently, when it was discovered that relatively good speech recognition was possible from the sum of a few temporal envelopes derived from band-pass filters spanning the range of speech-relevant frequencies (Shannon *et al.*, 1995), cochlear implant manufacturers migrated to using that more straight-forward approach.

Currently most cochlear implant processors operate by dividing the speech signal into a sequence of spectral channels, recovering the temporal envelope from each of those channels, and presenting those envelopes as time-varying amplitude signals to the corresponding electrodes. In general, channel boundaries are determined by dividing the range of band-pass frequencies being processed into equal intervals according to the distance covered along the normal basilar membrane. Nonetheless, current implants are unable to maximize channel capacity. Although there are typically close to twenty separate electrodes implanted within the cochlea, the limitations listed above mean there are far fewer independent channels of information actually available (e.g., Fishman *et al.*, 1997; Friesen *et al.*, 2001; Kiefer *et al.*, 2000; Shannon *et al.*, 1998), leading to spectral blurring of the signal. That blur diminishes speech recognition (e.g., Remez *et al.*, 2013).

In spite of the physiological limits imposed on the effective number of channels available to cochlear implant recipients, much of the experimentation examining ways to improve outcomes for users of these devices has involved manipulating the number of channels in either actual or simulated implant use. A major goal of implant design is to increase channel capacity. Only a handful of studies have investigated how the distribution of channels across the available frequency range influences outcomes. That lack of attention to the issue may have evolved because the first investigation into the question of channel distribution found little effect of channel boundaries on speech recognition (Shannon *et al.*, 1998). Nonetheless, a few studies have

[a]Author to whom correspondence should be addressed. Electronic mail: nittrouer.1@osu.edu

reported somewhat contradictory findings. For example, Fu and Shannon (2002) designed seven distribution schemes, varying from one that was exactly logarithmic in arrangement to one that was exactly linear. These varying strategies were used to map four separate channels in the devices of five listeners with cochlear implants. Results demonstrated that recognition of vowels was best for the logarithmic strategy and declined as the strategy approached a linear arrangement. Although the authors did not suggest it, one possible explanation for that outcome rests on the fact that vowel recognition depends strongly on listeners having well-defined spectral representations in the region of the voiced formants, especially the first and second formants. A logarithmic scale provides the most refined division of frequencies across the more apical implant channels where these relatively low-frequency formants would be represented. Another study (Fourakis *et al*., 2004) similarly observed better vowel recognition for cochlear implant users with maps that provided finer spectral resolution in the range of the first two formants. In both of those studies, however, advantages of the maps providing detailed formant structure were weaker or nonexistent for stimuli other than vowels. Nonetheless, these findings served to motivate interest in the possibility that finer-grained representation of the frequency region associated with vowel formant frequencies may facilitate better speech recognition for cochlear implant users. As a result, a primary goal of the current study was to test that hypothesis, that arranging available channels to represent as precisely as possible the first couple vowel formant frequencies facilitates speech recognition, when the number of channels of information is limited.

A second goal of the current study was to test the hypothesis that this consideration regarding channel assignment (of representing the first couple vowel formants as finely as possible) should facilitate speech recognition more strongly for children than adults. Evidence has been available for some time demonstrating that children, even infants, are sensitive to correspondences between acoustic speech structure and articulation. For example, Kuhl and Meltzoff (1982) reported that infants at 18 to 20 weeks of age looked longer at a face producing mouth gestures affiliated with a heard vowel than a face producing mouth gestures affiliated with a different vowel. Moreover, a third of the infants participating in that experiment clearly attempted to imitate the speech sounds presented to them. The authors concluded that both the detection of auditory-visual correspondence in speech signals and the vocal imitation reflected the infants' sensitivity to the relationship between audition and articulation. For deaf infants and children, exploiting this sensitivity by placing channel boundaries to represent formant structure could facilitate the development of speech perception and production. The benefit of optimal channel placement could be greater for children than for adults because children might not as effectively make use of other available clues to recognition, such as syntactic or semantic context, that can compensate when speech-relevant structure is not well represented. In the early years of life, it may be that sensitivity to correspondences between the acoustic speech signal and articulation is especially important, because it provides the initial structure for linguistic form. In turn, having those early (articulatory) representations allows children to discover other kinds of linguistic structure (e.g., syntax and semantics), which come to be used later to facilitate speech recognition. In other words, children likely need those early articulatory representations in order to develop the lexical and grammatical knowledge that serves adults well when they have access to only degraded sensory inputs.

The first study to investigate children's speech recognition using spectrally degraded signals providing only a few channels of amplitude structure was done by Eisenberg *et al.* (2000). In that study, the number of channels used in noise-vocoded signals varied between four and 32. Outcomes showed that 5- to 7-year-old children required more channels than adults to achieve similar levels of recognition, but listeners of all ages reached asymptote at eight channels, matching outcomes of other studies (e.g., Loizou *et al*., 1999). Signals in the Eisenberg *et al*. study consisted of the pass band between 300 and 6000 Hz. Boundaries between adjacent channels in the four-channel condition were 722, 1528, and 3066 Hz. For the six-channel condition, these boundaries were 550, 936, 1528, 2440, and 3842 Hz. In both conditions these boundaries created channels of equal intervals along the basilar membrane (Greenwood, 1990), but they did not provide channel divisions that were equally reasonable in how they represented important aspects of articulation. In the four-channel condition, the first formant would have been represented in the first and second channels, with some energy from the second formant also in the second channel. Vowels with higher second-formant frequencies would have been represented in the third channel, along with energy from some third formants. Having channels inconsistently represent formant frequencies in this manner could create uncertainty. For example, would a strong second channel specify a high first formant (associated with an open vowel) or a low second formant (associated with a back vowel)? In the six-channel condition, the first two channels would have neatly represented the first formant: Close vowels would have most of their first-formant energy in the first channel, and open vowels would have most of their energy in the second channel. Energy from the second formant would not have been present in that second channel. Instead, the third and fourth channels would have consistently represented vocalic second formants: Back vowels would have had their energy primarily in the third channel and front vowels would have had their energy primarily in the fourth channel. It is unlikely that any energy from the third formant would have been present in this fourth channel. Given differences in how channels aligned with vowel formant frequencies, the question could be asked of whether the outcomes of the Eisenberg *et al*. (2000) study might be more appropriately attributed to the placement of boundaries, rather than to sheer number of channels. That question was addressed in the current study. While listeners of all ages may have benefited from the channel structure in the six-channel stimuli, it is hypothesized that it is especially important for young children to have clearly represented formants.

The third and last goal of the current study was to examine the hypothesis that it would be useful to explicitly

preserve the time-varying structure of the speech signal, especially for children. This hypothesis is derived from work in the 1980s demonstrating a role of time-varying spectral structure in speech recognition (e.g., Kewley-Port *et al.*, 1983). In fact, in a presentation to the New York Academy of Sciences describing limits that should be imposed on alternative representations of speech signals for cochlear implants, Studdert-Kennedy (1983) based much of his recommendation on demonstrations that human listeners can accurately repeat sentences presented as only three time-varying sine waves replicating the first three formants (Remez *et al.*, 1981). This phenomenon had only just been discovered at the time Studdert-Kennedy made his presentation to the Academy, and its significance was not yet fully understood. It was Studdert-Kennedy who found meaning in the outcome, interpreting it with the statement "…what is crucial here is that the information preserved is not simply some trace of the formant structure, sufficient to specify instantaneous cavity relations, but the temporal patterns of spectral change that specify the forces controlling the movements by which cavity shapes are determined" (p. 36). That statement motivates the suggestion that finding ways to provide clear representations of time-varying formant structure could improve speech recognition through cochlear implants. Current processing schemes for cochlear implants, or noise vocoding in simulation studies, provide some representation of formant movement, but it is coarse. There may be other, speech-specific methods of processing signals to represent that structure in a more detailed manner. In studies with normal-hearing listeners using reduced spectral channels, detailed representations of formant movement are provided when sine waves replace formant frequencies.

Although it can be challenging to compare outcomes of experiments using noise-vocoded stimuli and those using sine-wave replicas, several reports provide evidence that children are more negatively affected in their speech perception when the time-varying information afforded by sine wave signals is missing. Nittrouer *et al.* (2009) presented four- and eight-channel noise-vocoded sentences to adults and 7-year-old children. Listeners of both ages performed better with the eight-channel than with the four-channel stimuli, but only the children performed better with the sine-wave stimuli than with the four-channel stimuli. That finding was replicated in a later study, using slightly different sentences and testing younger children (Nittrouer and Lowenstein, 2010). These outcomes across studies suggest that children could benefit even more than adults from speech-based processing strategies that clearly represent the time-varying structure of formant frequencies; those kinds of strategies have not been strongly pursued in implant designs since formant extraction methods were discontinued. However, the studies of Nittrouer *et al.* (2009) and Nittrouer and Lowenstein (2010) did not place channels in the noise-vocoded stimuli in such a way as to maximize the manner in which vowel formant frequencies were represented. Consequently, it is possible that the more precise representation of vowel formant frequencies, rather than the time-varying nature of that representation, accounted for the benefits shown by children in those earlier studies. That question was addressed by the current study.

II. EXPERIMENT 1

This first experiment was undertaken to examine whether the placement of channel boundaries affects speech recognition. It was hypothesized that placing boundaries to divide the acoustic space associated with each of the first and second formants into two well-defined channels should lead to better recognition. It was further hypothesized that the benefit of placing channels to optimize how well the signal specifies formant structure should be greater for children than for adults.

Although not primary aims of the study, two additional questions were addressed in this experiment. First, it was asked if the benefits of having channels that carefully divide in half the acoustic space of the first and second formants would facilitate vowel recognition more than consonant recognition. That outcome might be expected, given that channel division was based specifically on preserving vowel structure. It was also predicted by the outcomes of studies by Fu and Shannon (2002) and Fourakis *et al.* (2004). However, differences in design across studies made it an interesting question to explore further.

Second, the question was asked of whether having channel divisions specifically designed to preserve formant structure would facilitate the recognition of words in sentences more than words presented in isolation. That outcome would be predicted because the speech-based channel division would allow some—albeit coarse—representation of the time-varying structure of the first and second formants, in the form of the relative amplitudes across channels. The ability to track movement in formant frequencies over signal stretches longer than a word should aid speech recognition when signals are spectrally degraded. Listeners should be able to resolve any uncertainty that may exist in brief samples by following the formant trajectory.

A. Method

1. Participants

Sixty-three listeners participated in this experiment: 20 adults between the ages of 18 and 37, twenty-one 7-year-olds (ranging from 6 years, 11 months to 7 years, 5 months) and twenty-two 5-year-olds (ranging from 5 years, 1 month to 5 years, 10 months). All listeners were native speakers of American English, and all passed hearing screenings at 20 dB hearing level for the frequencies 0.5, 1, 2, 4, and 6 kHz. The children were all free from significant histories of otitis media, defined as more than six episodes before the age of 3 years old.

To ensure that all participants had language abilities within the normal range, they were screened. Adults were given the reading subtest of the Wide Range Achievement Test 4 (WRAT; Wilkinson and Robertson, 2006) and needed to demonstrate at least an 11th-grade reading level to participate. Children were given the Goldman Fristoe 2 Test of

Articulation (Goldman and Fristoe, 2000) and needed to score at or better than the 30th percentile for their age.

All listeners were also given the Expressive One-Word Picture Vocabulary Test—4th Edition (EOWPVT; Martin and Brownell, 2011) and were required to achieve a standard score of at least 90 (25th percentile). The mean EOWPVT standard score for adults was 105 (SD = 10), which corresponds to the 63rd percentile. The mean EOWPVT standard score for both 7- and 5-year-olds was 114 (SD = 12), which corresponds to the 82nd percentile. These scores indicate that the adult listeners had expressive vocabularies that were slightly above the mean of the normative sample used by the authors of the EOWPVT and that the children had expressive vocabularies closer to 1 SD above the normative mean.

2. Equipment

All materials for presentation were recorded in a sound booth, directly onto the computer hard drive, via an AKG C535 EB microphone, a Shure M268 amplifier, and a Creative Laboratories Soundblaster soundcard. Listening tests took place in a sound booth, with the computer that controlled the experiment in an adjacent room. Stimuli were stored on a computer and presented through a Samson headphone amplifier and AKG-K141 headphones. The hearing screening was done with a Welch Allyn TM262 audiometer and TDH-39 headphones. All test sessions were video-recorded using a Sony HDR-XR550V video recorder so that scoring could be done later. Participants wore Sony FM microphones that transmitted speech signals to an ARTcessories PowerMix III mixer. The speech signals, along with the stimuli, were transmitted from the mixer directly into the line input of the camera. This ensured good sound quality for all recordings.

3. Stimuli

The stimuli for this experiment consisted of five-word sentences that were highly meaningful, four-word sentences that were syntactically correct but not meaningful, and phonetically balanced CVC word lists. Two kinds of sentence materials were used to ensure that any effects that were observed—especially those related to listener age—could not be attributed to how meaningful those sentences were (i.e., semantic context effects). Words and sentences were used in order to test the hypothesis that any observed benefits of having specifically speech-based channel divisions might be greater for words in sentences where time-varying formant structure across word boundaries could be facilitative.

Fifty-four of the 72 five-word sentences (four for practice, 50 for testing) used in Nittrouer and Lowenstein (2010) were used. These sentences are syntactically correct and semantically predictable, and follow a subject-predicate structure (e.g., *Flowers grow in the garden*). They originally came from the Hearing in Noise Test (Nilsson *et al.*, 1994). In addition, 54 of the 56 four-word sentences (4 for practice, 50 for testing) used in Nittrouer *et al.* (2014) were presented in this study. These sentences are comprised entirely of monosyllabic content words and are syntactically correct but semantically anomalous (e.g., *Ducks teach sore camps*). Finally, 19 word lists (18 for testing, 1 for practice) from Mackersie *et al.* (2001) were used as well. Each word list consisted of 10 phonetically balanced CVC words. All of the sentences and words were recorded by an adult male talker of American English at a 44.1-kHz sampling rate with 16 bit digitization.

To create the vocoded stimuli, the same MATLAB routine was used as in previous experiments (e.g., Nittrouer and Lowenstein, 2010; Nittrouer *et al.*, 2009; Nittrouer *et al.*, 2014). All sentences were first band-pass filtered with a low-frequency cut-off of 250 Hz and a high-frequency cut-off of 8000 Hz. Next, each sentence was processed in each of two ways, using different channel divisions. Both schemes involved five channels, but the placement of channel boundaries differed. In the first scheme, boundaries were set as they were for the four-band condition in Eisenberg *et al.* (2000). Because 6000 Hz was the high-frequency cut-off in that earlier study, the fifth channel in the current stimuli extended from 6000 to 8000 Hz. These stimuli are termed the *standard* stimuli in this experiment. In this condition, the cutoff frequencies between channels for vocoding were 722, 1528, 3066, and 6000 Hz.

The second scheme used for placing channel boundaries was derived from the six-band condition of Eisenberg *et al.* (2000), but the spectrum from the upper cut-off of the fourth channel in that configuration to the high-frequency cut-off for stimuli in this study formed one broad channel. This scheme was designed to maximize frequency resolution in the lower frequency range, specifically for the first two formants. To achieve that goal, two channels were placed in the typical range of each of the first and second formants (F1 and F2). The goal behind this design was to provide two well-defined channels for the range of possible frequencies for each of those formants, and only for the range of each formant. Accordingly, the lowest two channels would specify vowel height (associated with F1) in a binary manner (relatively close or open) and the third and fourth channels would specify vowel fronting (associated with F2) in a binary manner (relatively front or back). These stimuli were termed the *speech-based* stimuli, and the cutoff frequencies between channels for vocoding were 550, 936, 1528, and 2440 Hz. Measurements made by Hillenbrand *et al.* (1995) supported the characterization of formant representation offered here. In that acoustic study, mean F1 frequency for male talkers for the most open vowel [ɑ] was measured as 768 Hz. Consequently, all F1 frequencies should be near or below that value, and within the range of the first two channels (i.e., between 250 and 936 Hz). According to Hillenbrand *et al.*, the mean F2 frequency of male speakers for the most backed vowel [u] is 997 Hz, and the mean F2 frequency for the most fronted vowel [i] is 2322 Hz. Consequently, all F2 frequencies should be between those values, and should fall within the range of the third and fourth channels (i.e., between 936 and 2440 Hz). With the exception of [ɹ]-colored vowels, all vocalic F3 frequencies should be above the cut-off of the fourth channel (i.e., 2440 Hz), as well as all spectral structure affiliated with fricative noises and release bursts.

All filtering used in the generation of these stimuli was done with digital filters that had greater than 50-dB

attenuation in stop bands, and had 1-Hz transition bands between pass- and stop-bands. Each channel was half-wave rectified and filtered below 20 Hz to remove fine structure. The temporal envelopes derived for separate channels were subsequently used to modulate white noise, limited to the same channels as those used to divide the speech signal. The resulting bands of amplitude-modulated noise were combined with the same relative amplitudes across channels as measured in the original speech signals. Root-mean-square amplitude was equalized across all stimuli.

4. Procedures

All procedures were approved by the Institutional Review Board of the Ohio State University. After participants (or their parents, in the case of children) signed the consent form, the hearing screening was administered.

Stimuli for testing were presented under headphones at 68 dB sound pressure level. There were four presentation blocks in the experiment based on materials: Five-word sentences (50), four-word sentences (50), and words (two blocks of nine lists each). The 18 word lists were divided into two shorter blocks in order to make the timing of all blocks as consistent as possible. Equal numbers of stimuli of each processing type (standard and speech-based) were presented in each block. For the sentence materials, standard and speech-based stimuli were mixed in each block, with the rule that no more than two standard or two speech-based stimuli could be presented in a row. For the word materials, standard and speech-based stimuli alternated between ten-word lists. Thus each sentence block consisted of 50 sentences (25 standard and 25 speech-based) and each word block consisted of nine word lists (either four standard and five speech-based for the first block and five standard and four speech-based for the second block, or vice versa).

There were four possible orders of presentation: (1) five-word sentences, half the word lists, four-word sentences, half the word lists; (2) four-word sentences, half the word lists, five-word sentences, half the word lists; (3) half the word lists, five-word sentences, half the word lists, four-word sentences; and (4) half the word lists, four-word sentences, half the word lists, five-word sentences. Equal numbers of adult and 7-year-old listeners completed each order of presentation. Five-year-olds listened to only five-word sentence stimuli, because pilot testing revealed that they had a difficult time paying attention to the four-word sentences and word lists.

Each type of material was preceded by a set of training stimuli. For blocks consisting of sentences, the listener heard and repeated a set of four sentences. For each practice sentence, the unprocessed version was played first and the listener was asked to repeat it. Then the processed version was presented and the listener was asked to repeat it. Two practice sentences were presented as standard stimuli, and two sentences were presented as speech-based stimuli. For the first block consisting of word lists, the listener heard and repeated one ten-word list. For each word, the unprocessed version was played first and the listener was asked to repeat it. Then the processed version was presented and the listener was asked to repeat it. Five words each were presented as standard and speech-based stimuli. The second block consisting of word lists was not preceded by training stimuli. Instead, the listener was instructed that he or she would be hearing and repeating words again.

During testing, listeners were seated across the table from the experimenter, with the video camera facing the listener. Listeners wore the FM transmitter, and all responses were video and audio recorded. Each sentence or word was played once, and the listener repeated what was heard. Children moved a game piece along a four-space game board after each block. This procedure provided a visible indicator of progress.

After testing was completed with the four blocks of test stimuli (one block for 5-year-olds), the two screening tasks were administered: WRAT and EOWPVT for adults, and the Goldman-Fristoe and EOWPVT for 7- and 5-year-olds.

5. Scoring and Analyses

Dependent measures were the percentages of words recognized correctly, both when presented in sentences and in isolation. Listeners had to respond with at least 10% correct words with at least one processing type for both the five-word sentence and four-word sentence materials (or in the case of 5-year-olds, the five-word sentence materials only) to have their data included in the study. This was done to avoid floor effects.

All responses were scored by the second author. In addition, the third author scored 25% of listener responses. Pearson product-moment correlation coefficients were computed between scores of the second author and the third author as a measure of inter-rater reliability. This procedure was done for data from 5-year-olds, 7-year-olds, and adults separately.

Although word recognition was the measure of primary interest, the numbers of whole sentences recognized correctly were also scored in order to quantify linguistic context effects on the recognition of words in sentences. For that purpose, j factors were computed using the formula described by Boothroyd and Nittrouer (1988):

$$j = \log(p_s)/\log(p_w), \quad (1)$$

where p_s is the probability of correct recognition of whole sentences, p_w is the probability of correct recognition of separate words, and j is the number of independent channels of information. Although j is typically between 1 and the number of words in the sentence, this value is not appropriately viewed as a number of actual words. Rather, j is a dimensionless factor that serves as an index of how strongly sentence context influences recognition. The smaller j is, the greater the effect of sentence context on recognition. One constraint in computing j factors is that this value becomes unstable when either word or sentence recognition scores are below 5% or greater than 95% correct. Thus, when scores are at those extremes, j should not be computed.

Data were screened for normal distributions and homogeneity of variance prior to conducting any statistical tests.

For inferential tests, arcsine transformations were used because data were given as percentages. A significance level of 0.05 was applied. Nonetheless, in reporting outcomes, precise significance levels are given when $p < 0.10$; for $p > 0.10$, outcomes are reported simply as not significant.

B. Results

One 7-year-old responded with less than 10% correct words with both types of processing for the four-word sentences, and two 5-year-olds responded with less than 10% correct words with both types of processing for the five-word sentences, so their data were not included. This resulted in data being included from twenty 7-year-olds and twenty 5-year-olds.

Inter-rater reliability was 0.99 across conditions for each age group. This was considered sufficiently reliable, and scores from the second author were used in analyses.

1. Five-word sentences

Thirty-one listeners had both word and sentence recognition scores between 5% and 95% correct for both processing conditions, so *j* factors were computed on their scores. Mean *j* factors for the three groups for each processing condition ranged from 2.5 to 2.9. A two-way, repeated-measures analysis of variance (ANOVA) was performed on these data, but neither age nor processing condition showed significant effects. Consequently it was concluded that context effects were similar across processing conditions and listener age, so could not explain any differences in recognition that might be observed.

Figure 1 shows mean recognition scores for both processing conditions, for each listener group. Listeners in all age groups achieved higher scores with speech-based processing. A two-way, repeated-measures ANOVA was performed on the recognition scores, with age as the between-subjects factor and processing as the repeated measure. The main effect of age was significant [F (2,57) = 106.68, $p < 0.001$, $\eta^2 = 0.789$], as were all *post hoc* contrasts among age groups ($p < 0.001$ for all contrasts using a Bonferroni adjustment for multiple contrasts). The main effect of processing was also significant [F (1,57) = 170.42, $p < 0.001$, $\eta^2 = 0.749$]. The age × processing interaction nearly reached significance [F (2,57) = 3.04, $p = 0.056$]. These results indicate that listeners were better at recognizing words for the speech-based stimuli than for the standard stimuli, and that recognition generally improved with increasing age. Although children may have benefited slightly more than adults from the enhanced processing strategy, that difference was not large enough to reach statistical significance. Mean difference scores (and SDs) between the speech-based and standard conditions were 13% (11%), 19% (10%), and 9% (6%), for 5-year-olds, 7-year-olds, and adults, respectively.

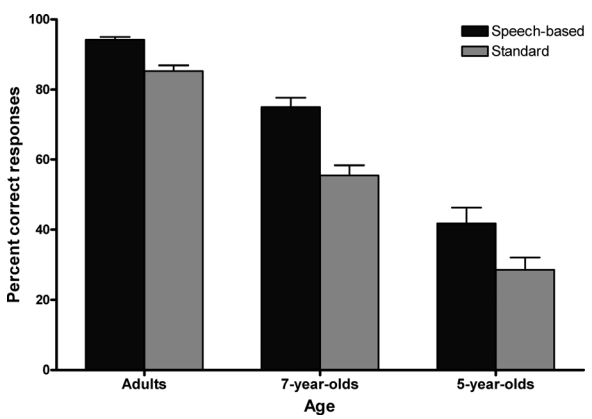

FIG. 1. Mean recognition scores for words in five-word sentences for adults, 7-year-olds, and 5-year-olds in experiment 1. Error bars are standard errors of the means.

2. Four-word sentences

Only two 7-year-olds had recognition scores for whole sentences that were better than 5% correct, and that was only with speech-based processing. Therefore it was not possible to compare *j* factors for adults and 7-year-olds. Nonetheless, given that no differences in context effects were observed across listener age with the five-word sentences, it seemed safe to conclude that the effect of syntax would be similar for these four-word sentences, as well, and earlier work with similar four-word sentences supported that conclusion (Nittrouer and Boothroyd, 1990). Children appear to use syntactic context effects in these simple sentences as well as adults. Semantics was not a factor because these sentences were anomalous.

Figure 2 shows mean recognition scores for the four-word sentences, for adults and 7-year-olds. As with the five-word sentence scores, both groups performed better with the speech-based stimuli than with the standard stimuli, and adults performed better than 7-year-olds. A two-way, repeated-measures ANOVA performed on these scores showed similar outcomes to those from the five-word sentences: The main effect of age was significant [F (1,38) = 63.73, $p < 0.001$, $\eta^2 = 0.626$], as was the main effect of processing

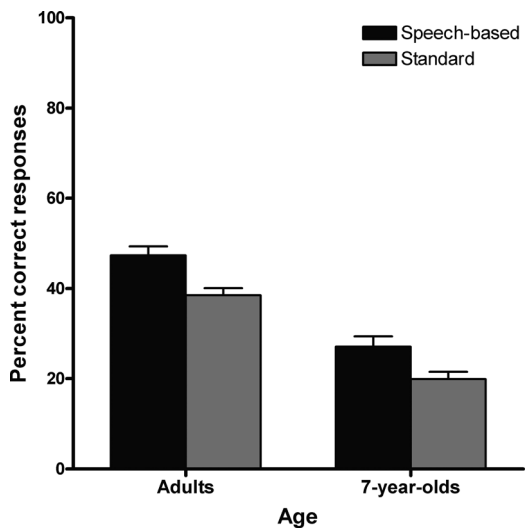

FIG. 2. Mean recognition scores for words in four-word sentences for adults and 7-year-olds in experiment 1. Error bars are standard errors of the means.

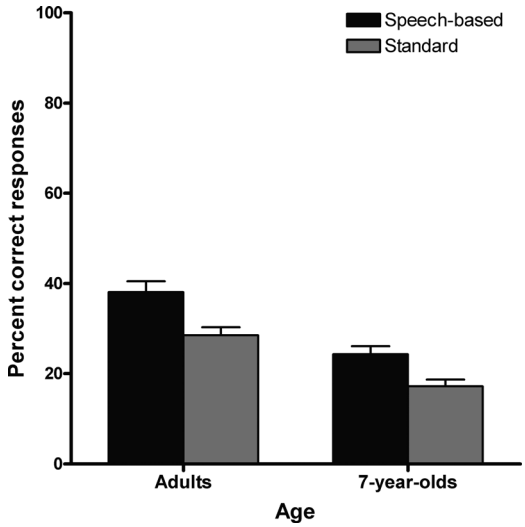

FIG. 3. Mean recognition scores for isolated words for adults and 7-year-olds in experiment 1. Error bars are standard errors of the means.

[F (1,38) = 41.23, $p < 0.001$, $\eta^2 = 0.520$]. However, the age × processing interaction was not significant.

3. Isolated words

Figure 3 shows mean recognition scores for words presented in isolation, for adults and 7-year-olds. Results appear to replicate those obtained with both kinds of sentence materials: Scores are better for the speech-based than for the standard stimuli, and better for adults than for 7-year-olds. A two-way, repeated-measures ANOVA performed on these scores showed similar outcomes to those from the sentence materials: The main effect of age was significant [F (1,38) = 25.4, $p < 0.001$, $\eta^2 = 0.401$], as was the main effect of processing [F (1,38) = 80.25, $p < 0.001$, $\eta^2 = 0.679$]. However, the age × processing interaction was not significant.

4. Is the advantage attributable to vowels?

In the speech-based stimuli, boundaries were placed to optimize the way that the spectral channels represented vowel F1 and F2. Consequently, it was predicted that vowel recognition would be especially facilitated by this processing strategy, and that prediction was tested. Figure 4 shows recognition scores for word initial consonant, vowel nucleus, and final consonant, for each processing condition, and for adults and 7-year-olds separately. It is clear that recognition improved in the speech-based condition, compared to the standard condition, for vowels more than for consonants. That outcome was observed for both groups of listeners. A three-way, repeated-measures ANOVA was performed on these data, with age as the between-subjects factor, and processing condition and segment as the repeated measures. All three main effects were found to be significant: age [F (1,38) = 23.31, $p < 0.001$, $\eta^2 = 0.380$]; processing [F (1,38) = 70.97, $p < 0.001$, $\eta^2 = 0.651$]; and segment [F (2,76) = 11.74, $p < 0.001$, $\eta^2 = 0.236$]. Only one two-way interaction was significant, and that was precisely the interaction being tested: processing × segment [F (2,76) = 36.01, $p < 0.001$, $\eta^2 = 0.487$]. This significant interaction reflects the finding that recognition for vowels was facilitated to a greater extent than for consonants when channels were established to optimize the representation for the first two formants.

One final outcome is notable: the three-way interaction of age × processing × segment was significant [F (2,76) = 5.18, $p = 0.008$, $\eta^2 = 0.120$], but the effect size was small. This interaction could be traced to a slightly larger effect of processing condition on vowel recognition for adults than for 7-year-olds.

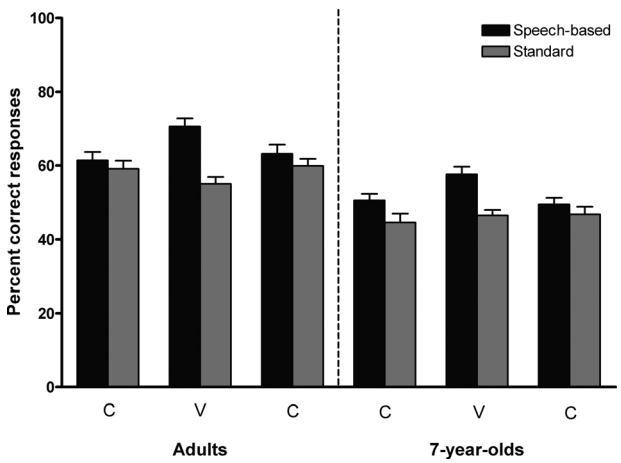

FIG. 4. Mean recognition scores for segments of isolated words in experiment 1. Error bars are standard errors of the means.

5. Is the advantage greater for words in sentences than for isolated words?

In continuous speech, coarticulatory effects cross syllable and word boundaries; in particular, formants transition smoothly among words. That continuity should help listeners discern formant frequencies, even when spectral structure is degraded. Therefore, it was hypothesized that the advantage of using a speech-based processing strategy would be greater for words presented in sentences where those transitions would be informative. Separate tests of this hypothesis were performed for five-word and four-word sentences, comparing scores of each to those for words in isolation. Specifically, three-way repeated-measures ANOVAs were performed on word recognition scores for each kind of sentence material, and words in isolation. Age was the between-subjects factor, and processing condition and materials were the repeated measures. The processing × materials interaction was the term from these analyses that would indicate whether the hypothesis was supported or not, and it was found to be significant for five-word sentences [F (1,38) = 20.71, $p < 0.001$, $\eta^2 = 0.353$], but not four-word sentences. Although the effect size was not large, this significant interaction reflects the finding that recognition for words in the five-word sentences was more affected by the processing strategy than recognition of the isolated words. Figure 5 illustrates this effect, by displaying mean recognition scores across adults and 7-year-olds, for each type of processing and materials.

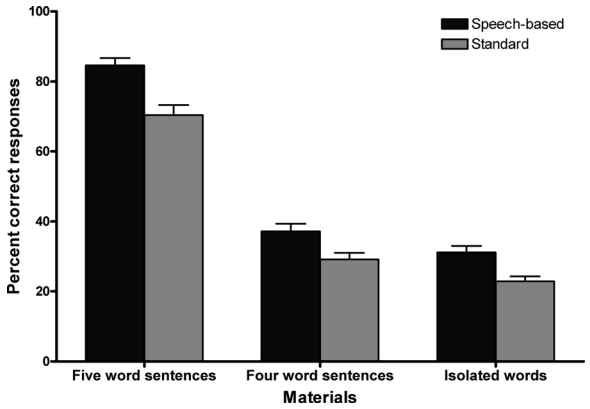

FIG. 5. Mean recognition scores for words in five-word sentences, words in four-word sentences, and isolated words, across adults and 7-year-olds in experiment 1. Error bars are standard errors of the means.

C. Discussion

This first experiment was undertaken to test two primary hypotheses: (1) Speech recognition would benefit from channels being placed to optimize the representation of the first two formants; and (2) The magnitude of this effect would be greater for children than for adults. Data collected to test these hypotheses revealed that the first hypothesis was supported, but the second was largely unsupported.

Two additional hypotheses were tested: (3) The effect of using a speech-based processing strategy would be greater for vowels than for consonants; and (4) The effect of using a speech-based processing strategy would be greater for words in sentences than for words in isolation. The first of these additional hypotheses was well-supported, as had been predicted due to the fact that the speech-based processing strategy was primarily designed to represent vowel formants. The second of these additional hypotheses was also supported, with recognition of words in the five-word sentences benefiting more from the speech-based processing than isolated words. However, the effect was not found for four-word sentences. That finding is probably because it was easier for the talker to maintain continuous production across word boundaries with the five-word than with the four-word sentences because the former were more natural than the latter. However, the possibility that the semantic context available for the five-word sentence materials evoked the greater recognition cannot be dismissed, and it is impossible based on these data to disambiguate the two possible accounts.

III. EXPERIMENT 2

This second experiment was undertaken to test the recommendation from Studdert-Kennedy (1983) that it would be especially important to preserve the time-varying structure of the speech signal in any representation provided through auditory prostheses. The rationale behind this recommendation is that the pattern of change across formants over time is itself informative; speech perception does not proceed by the listener taking a series of discrete "snapshots" of spectral structure. In this study it was further hypothesized that preserving the time-varying structure would be particularly important for children.

A. Method

1. Participants

Sixty-two listeners participated in this experiment: 20 adults between the ages of 18 and 28, twenty-two 7-year-olds (ranging from 6 years, 11 months to 7 years, 10 months) and twenty 5-year-olds (ranging from 5 years, 0 months to 5 years, 10 months). Although all listeners in this experiment were different from those of the first experiment, they met the same criteria as listeners in experiment 1.

The mean EOWPVT standard score for adults in this second experiment was 108 (SD = 13), which corresponds to the 70th percentile. The mean EOWPVT standard score for 7-year-olds was 117 (SD = 13) and for 5-year-olds it was 116 (SD = 12), which corresponded to the 87th and 86th percentiles, respectively. These scores indicate that adults had expressive vocabularies that were half of a SD above the mean of the normative sample, and children had expressive vocabularies roughly 1 SD above the normative mean.

2. Equipment

The same equipment was used as in experiment 1.

3. Stimuli

The stimuli for this experiment consisted of the same five-word and four-word sentences as used in experiment 1. In this experiment, those sentences were presented with speech-based processing and as sine-wave replicas. Isolated words were not included in this second experiment because any benefit of having time-varying spectral structure would likely extend primarily to connected speech.

The speech-based stimuli used in this experiment were the same as those used in experiment 1. Sine-wave stimuli were created with procedures used previously (Nittrouer and Lowenstein, 2010; Nittrouer *et al.*, 2009). A Praat script written by Darwin (2003) (available at http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/SWS) was used to create these stimuli. However, the formant tracks extracted with this script were adjusted until all of the first three formants closely matched the original speech signal, as represented in the spectrogram, and then the formant object was hand-adjusted as necessary to remove extraneous points. The script then generated a sine wave stimulus from the adjusted formant tracks. Root-mean-square amplitude was matched across all stimuli.

4. Procedures

In this experiment, speech-based and sine-wave stimuli were presented in separate blocks because the signals were qualitatively very different. Consequently, there were four blocks in the experiment (two kinds of materials × two processing conditions). The order of presentation was varied across listeners within each age group. As in experiment 1, 5-year-olds were presented with only five-word sentences.

Training and test procedures were otherwise the same as in experiment 1, as were scoring and analysis.

B. Results

Two 7-year-olds obtained scores of less than 10% correct for the four-word sentences with both types of processing, so their data were not included in these analyses. That left 20 participants in each age group.

1. Five-word sentences

Before analyzing the data for recognition scores, *j* factors obtained for each group were analyzed, as they had been in experiment 1. In this experiment, 30 listeners had both word and sentence recognition scores between 5% and 95% correct for both processing conditions. Mean *j* factors for the three groups for each processing condition ranged from 2.4 to 3.5. That is similar to what was found in the first experiment, although some values were a little higher. A two-way, repeated-measures ANOVA was performed on these data, but neither age nor processing condition showed significant effects. Consequently it was concluded that context effects were similar across processing conditions and listener age.

The next analysis that was done involved a series of *t* tests for each age group, comparing scores for listeners in this second experiment to those from listeners in the first experiment for the condition that was common across the two experiments: The speech-based stimuli. None of these tests was statistically significant, so it was concluded that the metrics were reliable.

Figure 6 shows mean recognition scores for both processing conditions, for each listener group. All age groups achieved higher scores for the sine-wave stimuli than for the speech-based stimuli. In this experiment, however, it appears that the effect of processing was greater for children than for adults. A two-way, repeated-measures ANOVA was performed on the recognition scores, with age as the between-subjects factor and processing as the repeated measure. The main effect of age was significant [F (2,57) = 126.34, $p < 0.001$, $\eta^2 = 0.816$], as were all *post hoc* contrasts among age groups ($p < 0.001$ for all contrasts using a Bonferroni adjustment for multiple contrasts). The main effect of processing was also significant [F (1,57) = 165.81, $p < 0.001$, $\eta^2 = 0.744$]. These results indicate that listeners were better at recognizing words with the sine-wave stimuli than with the speech-based stimuli, and recognition generally improved with increasing age. Finally, the age × processing interaction was significant [F (2,57) = 18.50, $p < 0.001$, $\eta^2 = 0.394$], indicating that the effect of processing decreased with increasing age. Mean difference scores (and SDs) between the sine-wave and speech-based conditions were 35% (17%), 17% (12%), and 4% (3%), for 5-year-olds, 7-year-olds, and adults, respectively.

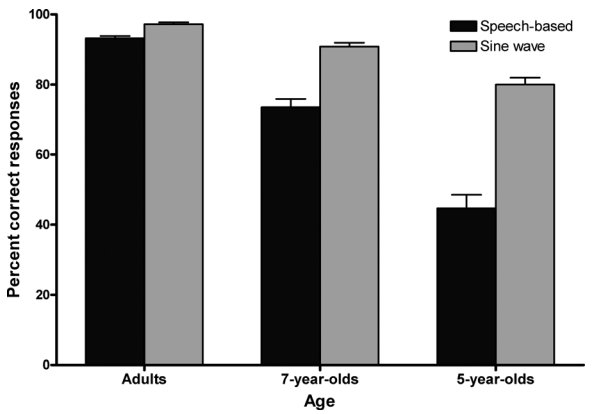

FIG. 6. Mean recognition scores for words in five-word sentences for adults, 7-year-olds, and 5-year-olds in experiment 2. Error bars are standard errors of the means.

2. Four-word sentences

Fourteen listeners (in the adult and 7-year-old groups) had word and whole-sentence recognition scores between 5% and 95% correct for these four-word sentences in both processing conditions, so *j* factors could be computed on their scores. A two-way ANOVA performed on these *j* factors failed to reveal either a significant main effect of age or of processing condition.

Scores were also compared across the two experiments for the speech-based stimuli, for adults and 7-year-olds separately. As with five-word sentences, no significant difference was observed for either group, indicating that these scores have adequate reliability.

Figure 7 shows mean word recognition scores for these four-word sentences, for each processing condition, and for adults and 7-year-olds separately. A two-way ANOVA performed on these data revealed significant main effects of age [F (1,38) = 35.31, $p < 0.001$, $\eta^2 = 0.482$]; and processing [F (1,38) = 16.24, $p < 0.001$, $\eta^2 = 0.299$]. However, for these materials, the age × processing interaction was not significant.

C. Discussion

This experiment was undertaken to examine whether time-varying spectral structure would provide any benefit to recognition of alternative representations of speech signals

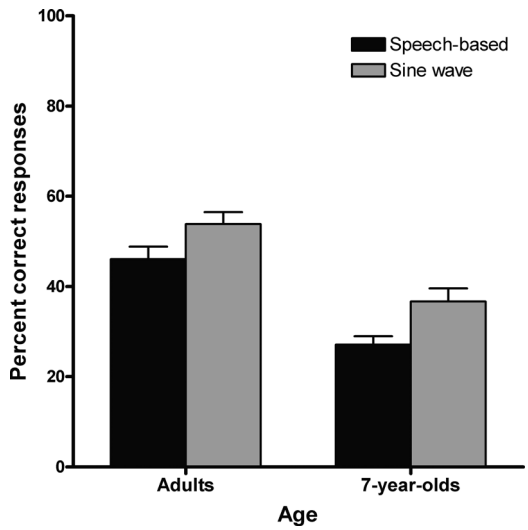

FIG. 7. Mean recognition scores for words in four-word sentences for adults and 7-year-olds in experiment 2. Error bars are standard errors of the means.

over what is provided by placing static spectral channels in a way that optimizes the representation of the first two formants. It was hypothesized that children would especially benefit from this kind of time-varying structure because earlier experiments had demonstrated that children performed disproportionately better than adults with sine-wave signals, compared to noise-vocoded signals (Nittrouer and Lowenstein, 2010; Nittrouer *et al.*, 2009). However, in that earlier work, channels in the noise-vocoded stimuli were not placed to optimize how well they represented formants, so it was not known before the current study was conducted if time-varying structure would be more facilitative than that placement strategy alone. Results of this second experiment supported the hypothesis that time-varying formant structure provides benefits over what is afforded, even by well-placed static channels, and that effect was stronger for children than for adults—at least for the five-word materials.

There was one confound in the design of stimuli across the two conditions in this second experiment that presented a possible alternative interpretation of the results. Stimuli in the speech-based condition were spectrally broad, whereas the sine wave stimuli provided spectrally narrow representations of formant structure. It might be argued that listeners—especially children—simply prefer narrower formant representations. However, outcomes of an earlier study (Lowenstein *et al.*, 2012) contradicts that interpretation. That study compared recognition of two sets of processed sentences, similar in design to noise-vocoded and sine-wave stimuli. In this case, however, both sets of stimuli consisted of 10-Hz wide noise bands. Results showed that even though both sets of stimuli were equivalent in width of the spectral bands, listeners performed better with the time-varying signals, and children disproportionately so. Thus, these results were able to demonstrate that it is explicitly the time-varying nature of sine wave stimuli that is critical to the advantage observed in children's speech recognition, rather than the fact that sine wave stimuli provide a narrower representation of spectral structure.

IV. GENERAL DISCUSSION

The overarching goal of this study was to examine whether there would be any reason to suggest that signal processing for cochlear implants might benefit from a speech-specific strategy. Although this approach has largely been abandoned, newer implant technologies—those developed since speech-specific strategies were being used—might permit better implementation of such methods. It was further hypothesized that these strategies might be more important to speech recognition for children than for adults, leading to the possibility that different processing algorithms could be used at different stages across the life span.

The first experiment in this study tested the hypothesis that speech recognition would be facilitated if static spectral channels were placed to optimize how well they represented the first two voiced formants. In this case, placing these channels to represent the first two formants meant that they essentially divided the acoustic space affiliated with each formant in half. As a result, vowel height (close or open) could be recovered from the first two channels, and front/back placement could be recovered from the third and fourth channels. In contrast to that strategy, channel boundaries for real and simulated processing are typically placed so that channels cover equivalent distances along the basilar membrane. When this latter approach is implemented in simulation studies, it is consistently observed that performance improves as the number of channels increases. The current study emerged from a desire to evaluate whether that finding might actually be attributed to the fact that, serendipitously, a better match between channel placement and formant representation is achieved when more channels are available.

Listeners in all three age groups performed better in the first experiment with channels placed to optimize the representation of the first two formants, thus supporting the suggestion that earlier findings of better recognition with more channels could at least partly be explained by more accurate representation of these important formants. Further support for that conclusion was gathered from the finding that it was specifically vowel recognition that gained the most from the speech-based processing strategy. Moreover, the magnitude of the improvement was similar to what was found by Eisenberg *et al.* (2000) when the number of channels increased from four to six: Close to a 20% improvement was observed by those authors, and that was precisely what was found in the first experiment of this current study.

The second hypothesis tested in the first experiment of the current study was that children would benefit more from the speech-based approach to channel placement. However, that hypothesis was not well-supported: The improvement observed when channel placement was speech-based was similar in magnitude across all listener groups, although it came close to showing a significant age effect for five-word sentences.

The second experiment was conducted to test the hypothesis that adding the time-varying aspect to the signal structure would improve recognition over and above what was achieved by placing static channels to optimize formant representation. In this case, a benefit was observed for listeners of all ages, so the hypothesis was supported. In addition, the effect was clearly greater for children than adults, for the five-word sentences.

A trend observed across both experiments was that the magnitude of benefit observed for either speech-based channel placement (experiment 1) or time-varying signal structure (experiment 2) was greater for the most strongly connected speech stimuli, five-word sentences. Although not significant in the first experiment, there was also a trend toward that advantage for strongly connected speech to be greatest for children than for adults. It is proposed that the benefit found for these stimuli arises from the fact that it is perceptually advantageous to be able to track formant frequencies across stretches of speech that are longer than a word, especially when the signal is spectrally degraded.

The clinical implications of the current study are that designs for implant processing should consider how formant structure is represented. This consideration was a focus of early implant design efforts, but is not especially a strong consideration at present. Instead, current strategies seek to

divide the pass band of frequencies across the set of available electrodes as equally as possible, which in the terminology of Levitt (1991) is a nonspeech-specific approach. There have been questions raised regarding whether this division is best obtained with a logarithmic or linear arrangement (e.g., Fu and Shannon, 2002), but those studies have not specifically considered articulatory properties of the signal. Furthermore, outcomes of the study reported here indicate that finding ways to preserve specifically time-varying aspects of formant structure could be useful, especially for children. Earlier processing strategies tried to provide this kind of time-varying structure, but it is less robustly represented with standard implant designs that utilize static channel allocations.

V. LIMITATIONS AND FUTURE DIRECTIONS

As encouraging as the current findings are, further empirical study is required before strategies designed to preserve formant structure could be effectively implemented in cochlear implants. In this study, noise-vocoded signals were presented to listeners with normal hearing. Consequently, the alignment between signal frequency and place on the basilar membrane was normal, a situation that cannot be assumed in impaired auditory systems with cochlear implants. In other studies—those unconcerned with the potential benefits of these speech-based processing strategies—it has been observed that outcomes are best when input frequency is mapped onto the cochlear location normally processed by that frequency (e.g., Başkent and Shannon, 2004). Because the frequencies affiliated with the first two formants are generally low, maintaining that normal frequency-place map may be impossible; implant insertion may not be deep enough. Thus, the question arises of how distorted the frequency-place map can be and still facilitate the benefits of precise representation of these first two formants. In addition, the output of a formant-tracking algorithm could encounter the same spectral-blurring effects (due to spread of excitation along the basilar membrane) that arise in other processing algorithms, although current-steering algorithms are being developed to try to evade this constraint (e.g., Donaldson *et al.*, 2011; Firszt *et al.*, 2007). Future investigation is warranted to examine these potential challenges to using speech-based processing strategies in cochlear implants.

ACKNOWLEDGMENTS

The authors wish to thank Jamie Kuess for writing the software to present stimuli. This work was supported by Grant No. R01 DC000633 from the National Institutes of Health, National Institute on Deafness and Other Communication Disorders.

Başkent, D., and Shannon, R. V. (**2004**). "Frequency-place compression and expansion in cochlear implant listeners," J. Acoust. Soc. Am. **116**, 3130–3140.

Blamey, P. J., Martin, L. F., and Clark, G. M. (**1985**). "A comparison of three speech coding strategies using an acoustic model of a cochlear implant," J. Acoust. Soc. Am. **77**, 209–217.

Boothroyd, A., and Nittrouer, S. (**1988**). "Mathematical treatment of context effects in phoneme and word recognition," J. Acoust. Soc. Am. **84**, 101–114.

Darwin, C. (**2003**). "Sine-wave speech produced automatically using a script for the PRAAT program," available at http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/SWS/ (Last viewed February 20, 2014).

Donaldson, G. S., Dawson, P. K., and Borden, L. Z. (**2011**). "Within-subjects comparison of the HiRes and Fidelity120 speech processing strategies: Speech perception and its relation to place-pitch sensitivity," Ear Hear. **32**, 238–250.

Eisenberg, L. S., Shannon, R. V., Schaefer Martinez, A., Wygonski, J., and Boothroyd, A. (**2000**). "Speech recognition with reduced spectral cues as a function of age," J. Acoust. Soc. Am. **107**, 2704–2710.

Falkner, A., Ball, V., Rosen, S., Moore, B. C. J., and Fourcin, A. (**1992**). "Speech pattern hearing aids for the profoundly impaired: Speech perception and auditory abilities," J. Acoust. Soc. Am. **91**, 2136–2155.

Firszt, J. B., Koch, D. B., Downing, M., and Litvak, L. (**2007**). "Current steering creates additional pitch percepts in adult cochlear implant recipients," Otol. Neurotol. **28**, 629–636.

Fishman, K. E., Shannon, R. V., and Slattery, W. H. (**1997**). "Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor," J. Speech Lang. Hear. Res. **40**, 1201–1215.

Fourakis, M. S., Hawks, J. W., Holden, L. K., Skinner, M. W., and Holden, T. A. (**2004**). "Effect of frequency boundary assignment on vowel recognition with the Nucleus 24 ACE speech coding strategy," J. Am. Acad. Audiol. **15**, 281–299.

Fourcin, A. J. (**1990**). "Prospects for speech pattern element aids," Acta Otolaryngol. Suppl. **469**, 257–267.

Friesen, L. M., 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., and Shannon, R. V. (**2002**). "Frequency mapping in cochlear implants," Ear Hear. **23**, 339–348.

Goldman, R., and Fristoe, M. (**2000**). *Goldman Fristoe 2: Test of Articulation* (American Guidance Service, Circle Pines, MN), 146 pp.

Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species–29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.

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

Kewley-Port, D., Pisoni, D. B., and Studdert-Kennedy, M. (**1983**). "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," J. Acoust. Soc. Am. **73**, 1779–1793.

Kiefer, J., von, I. C., Rupprecht, V., Hubner-Egner, J., and Knecht, R. (**2000**). "Optimized speech understanding with the continuous interleaved sampling speech coding strategy in patients with cochlear implants: Effect of variations in stimulation rate and number of channels," Ann. Otol. Rhinol. Laryngol. **109**, 1009–1020.

Kuhl, P. K., and Meltzoff, A. N. (**1982**). "The bimodal perception of speech in infancy," Science **218**, 1138–1141.

Levitt, H. (**1991**). "Signal processing for sensory aids: A unified view," Am. J. Otol. **12**(Suppl), 52–55.

Loizou, P. C., Dorman, M., and Tu, Z. (**1999**). "On the number of channels needed to understand speech," J. Acoust. Soc. Am. **106**, 2097–2103.

Lowenstein, J. H., Nittrouer, S., and Tarr, E. (**2012**). "Children weight dynamic spectral structure more than adults: Evidence from equivalent signals," J. Acoust. Soc. Am. **132**, EL443–EL449.

Mackersie, C. L., Boothroyd, A., and Minniear, D. (**2001**). "Evaluation of the Computer-Assisted Speech Perception Assessment Test (CASPA)," J. Am. Acad. Audiol. **12**, 390–396.

Martin, N., and Brownell, R. (**2011**). *Expressive One-Word Picture Vocabulary Test (EOWPVT)*, 4th ed. (Academic Therapy Publications, Novato, CA), 99 pp.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.

Nittrouer, S., and Boothroyd, A. (**1990**). "Context effects in phoneme and word recognition by young children and older adults," J. Acoust. Soc. Am. **87**, 2705–2715.

Nittrouer, S., and Lowenstein, J. H. (**2010**). "Learning to perceptually organize speech signals in native fashion," J. Acoust. Soc. Am. **127**, 1624–1635.

Nittrouer, S., Lowenstein, J. H., and Packer, R. (**2009**). "Children discover the spectral skeletons in their native language before the amplitude envelopes," J. Exp. Psychol. Hum. Percep. Perform. **35**, 1245–1253.
Nittrouer, S., Tarr, E., Bolster, V., Caldwell-Tarr, A., Moberly, A. C., and Lowenstein, J. H. (**2014**). "Low-frequency signals support perceptual organization of implant-simulated speech for adults and children," Int. J. Audiol. **53**, 270–284.
Remez, R. E., Cheimets, C. B., and Thomas, E. F. (**2013**). "On the tolerance of spectral blur in the perception of words," Proc. Meet. Acoust. **19**, 1–6.
Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–949.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Shannon, R. V., Zeng, F. G., and Wygonski, J. (**1998**). "Speech recognition with altered spectral distribution of envelope cues," J. Acoust. Soc. Am. **104**, 2467–2476.
Studdert-Kennedy, M. (**1983**). "Limits on alternative auditory representations of speech," Ann. N.Y. Acad. Sci. **405**, 33–38.
Tye-Murray, N., Lowder, M., and Tyler, R. S. (**1990**). "Comparison of the F0F2 and F0F1F2 processing strategies for the Cochlear Corporation cochlear implant," Ear Hear. **11**, 195–200.
Wilkinson, G. S., and Robertson, G. J. (**2006**). *The Wide Range Achievement Test (WRAT)*, 4th ed. (Psychological Assessment Resources, Lutz, FL), 494 pp.
Wilson, B. S., Finley, C. C., Lawson, D. T., Wolford, R. D., Eddington, D. K., and Rabinowitz, W. M. (**1991**). "Better speech recognition with cochlear implants," Nature **352**, 236–238.