Combined spectral and temporal enhancement to improve cochlear-implant speech perception

Aparajita Bhattacharya^{a)}

Center for Hearing Research, Department of Biomedical Engineering, University of California, Irvine, California 92697

Andrew Vandali

The HEARing Cooperative Research Centre, Carlton, 3053, Victoria, Australia

Fan-Gang Zeng

Center for Hearing Research, Departments of Anatomy and Neurobiology, Biomedical Engineering, Cognitive Sciences, and Otolaryngology–Head and Neck Surgery, University of California, Irvine, California 92697

(Received 28 September 2009; revised 21 August 2011; accepted 23 August 2011)

The present study examined the effect of combined spectral and temporal enhancement on speech recognition by cochlear-implant (CI) users in quiet and in noise. The spectral enhancement was achieved by expanding the short-term Fourier amplitudes in the input signal. Additionally, a variation of the Transient Emphasis Spectral Maxima (TESM) strategy was applied to enhance the short-duration consonant cues that are otherwise suppressed when processed with spectral expansion. Nine CI users were tested on phoneme recognition tasks and ten CI users were tested on sentence recognition tasks both in quiet and in steady, speech-spectrum-shaped noise. Vowel and consonant recognition in noise were significantly improved with spectral expansion combined with TESM. Sentence recognition improved with both spectral expansion and spectral expansion combined with TESM. The amount of improvement varied with individual CI users. Overall the present results suggest that customized processing is needed to optimize performance according to not only individual users but also listening conditions.

© 2011 Acoustical Society of America. [DOI: 10.1121/1.3641401]

PACS number(s): 43.66.Ts, 43.71.Ky, 43.71.Es, 43.71.Gv [BCM]

Pages: 2951–2960

I. INTRODUCTION

Although many cochlear-implant (CI) users enjoy reasonable performance in quiet, perception in a noisy environment still remains a challenge. In general, CI users are much more susceptible to noise than normal-hearing (NH) listeners (Dorman *et al.*, 1998; Hochberg *et al.*, 1992; Nelson *et al.*, 2003; Stickney *et al.*, 2004; Zeng and Galvin, 1999; Zeng *et al.*, 2005). To understand 50% of speech in noise, CI users typically need a 10–15 dB higher signal-to-noise ratio (SNR) than NH listeners in steady noise, and as much as 30 dB in fluctuating background noise (Nelson *et al.*, 2003; Zeng *et al.*, 2005).

CI speech recognition in noise is adversely affected by reduced internal representation of spectral contrast compared to NH listeners. One reason for this is poor spectral resolution due to the broad spread of neural excitation from electrical stimulation in the cochlea (Chatterjee and Shannon, 1998; Cohen, 2009; Cohen *et al.*, 2003). This broad electrical stimulation may be further complicated by poor neural survival and suboptimal electrode placement, thereby reducing the number of independent spectral channels and distorting the tonotopic organization in CI users (Busby *et al.*, 1993; Cohen *et al.*, 1996). Other factors that could distort the internal representation of spectral contrast are reduced

number of discriminable intensities and narrow electric dynamic range (Loizou *et al.*, 2000; Nelson *et al.*, 1996). Finally, when listening to speech in noise, spectral contrast is reduced by noise, which fills the valleys of the speech spectral envelope. As a result, CI users typically need more spectral and temporal contrast than NH listeners to understand speech in noise (Loizou and Poroy, 2001). Here we hypothesize that front end spectral and temporal enhancement techniques to compensate for the effect of reduced internal representation of spectral contrast may benefit CI users in noise.

Several spectral enhancement techniques have been tested using NH and hearing-impaired (HI) listeners in attempts to improve their speech understanding in noise (Baer *et al.*, 1993; Bunnell, 1990; Clarkson and Bahgat, 1991; Franck *et al.*, 1999; Lyzenga *et al.*, 2002). In particular, Lyzenga *et al.* (2002) used a spectral expansion scheme to enhance parts of the spectrum corresponding to vowel formants. The spectral expansion was achieved by dividing the signal into overlapping short frames, and expanding the amplitude spectrum of each frame. Additionally, the upward-spread-of-masking was compensated by applying linearly increasing gain from 0 to 6 dB over the frequency range spanned by the first three formants of vowels in speech. The root-mean-square (rms) value of the expanded signal was set equal to its original value on a frame-by-frame basis. As a result, the temporal contrast between the frames was preserved. NH listeners listening to spectrally smeared speech

^{a)}Author to whom correspondence should be addressed. Electronic mail: abhattac@uci.edu

showed improvement for conditions in which spectral expansion and high-frequency emphasis were combined.

The spectral enhancement technique employed in this study is similar to the spectral expansion scheme of [Lyzenga et al. \(2002\)](#). However, the present study did not normalize the rms value of the expanded signal with respect to its original value on a frame-by-frame basis. This difference introduced a potential increase in temporal contrast between frames. In a sentence corrupted with noise, the noise may fill in the gaps between speech segments in such a way that words/syllables are less delineated in time. Increasing the contrast between temporal frames can potentially suppress the level of lower-intensity noise that intervenes between words/syllables in the sentences, and thereby improve the intelligibility of speech.

Consonants are generally weaker than vowels, and hence increasing the temporal contrast between the vowel and consonant frames will decrease the consonant-to-vowel intensity ratio. This may reduce the intelligibility of some consonants and syllables. A plausible method for counterbalancing the negative effect of spectral expansion on the consonant-vowel intensity ratio is to enhance the relatively weak short-duration consonant cues before expansion. The transient emphasis spectral maxima (TESM) strategy was proposed to enhance such short-duration transient cues in speech, e.g., the burst of noise accompanying a stop ([Vandali, 2001](#)). Speech was analyzed by a bank of 16 band-pass filters and additional gain was applied to a band whenever there was a rapid rise in its envelope. The gain applied was higher when there was a rapid rise followed by a rapid fall (e.g., as might occur for a consonant burst) than when there was a rapid rise followed by a relatively constant level (e.g., as occurs at the onset of a vowel). No gain was applied

for steady or falling envelope levels. The study showed that the strategy improved the perception of nasal, stop, and fricative consonants. We therefore hypothesized that enhancing the transient short-duration cues with TESH before spectral expansion would compensate for the decrease in consonant-vowel intensity ratio due to expansion and improve the perception of some consonants and syllables.

The present study was aimed at assessing the effect of combining spectral expansion and TESH on the speech perception of CI users in quiet and in noise. Three types of processing, including expansion (E), TESH, and TESH combined with expansion (TESM_E), were evaluated as preprocessing strategies to cochlear implant processing. CI users were tested on vowel, consonant, and sentence recognition tests in quiet, and in steady, speech-spectrum-shaped noise (SSN).

II. METHODS

A. Processing

Figure 1 shows schematically the strategies evaluated in this study. The strategies were implemented offline using MATLAB (The Mathworks, Natick, MA) to preprocess the test speech material for subsequent perceptual testing. The test material was sampled at 32 kHz for vowels, 44.1 kHz for consonants and 20 kHz for sentences.

For spectral expansion (E), consecutive segments of 32 ms (vowels), 23.3 ms (consonants), and 25.6 ms (sentences) durations were windowed using a Hanning window with 75% overlap. Each frame was passed through a pre-emphasis filter, a first-order high-pass filter with cutoff at 500 Hz. Next, a fast Fourier transform (FFT) was performed and the amplitude spectrum was raised to the power 1.6. The

FIG. 1. Block diagram of the processing schemes. (A) PE—pre-emphasis, DE—de-emphasis. (B) Slow varying envelopes E_c —current window, E_p —past window, E_f —future window, and $k = 2.5$.

value of the exponent was chosen as a tradeoff between the degree of spectral contrast enhancement and distortion. An inverse Fourier transform (IFFT) was applied to the modified amplitude spectra and the original phase spectra followed by de-emphasis filtering to create the processed segments. The segments were again windowed with a Hanning window and added with 75% overlap. The rms level of the entire processed stimulus was normalized with respect to the rms level of the original stimulus, rather than on a frame-by-frame basis as in earlier studies. The differences between the processing used in the current study and that of [Lyzenga et al. \(2002\)](#) are: (1) A pre-emphasis filter was used in the current study to flatten the long-term average speech spectrum of the material. (2) Spectral expansion was applied throughout the spectrum in the current study, whereas [Lyzenga et al. \(2002\)](#) applied expansion only to the frequency region where the speech formants occurred (400–4 500 Hz). (3) The rms value of the processed signal was not normalized with respect to the original speech on a frame-by-frame basis in the current study, whereas it was in [Lyzenga et al. \(2002\)](#).

For TESM, consecutive segments of 4 ms (vowels), 2.9 ms (consonants), and 3.2 ms (sentences) durations were windowed using a Hanning window with 75% overlap. Next, an FFT was performed and, for each FFT bin, the slow-varying envelope was estimated by averaging the bin amplitudes over a 20 ms duration. Three consecutive averages separated by 20 ms were maintained, representing the past (E_p), current (E_c), and future (E_f) slow-varying envelope levels. The amplitude of each bin (e_n), delayed by 30 ms (i.e., so as to be relative to the midpoint of E_c), was then modified according to the following expression:

$$e_n(\text{mod}) = e_n \left(1 + k \frac{2E_c - 2E_p - E_f}{E_c + E_p + E_f} \right). \quad (1)$$

The value of k was chosen to be 2.5 [as opposed to 2.0, as used by [Vandali \(2001\)](#)] so that the consonant cues (spanning a time period of no longer than ~ 60 ms) were not only preserved but also amplified when TESM was combined with E. An IFFT was performed on the modified amplitudes and original phases. The frames were windowed using a Hanning window and added with 75% overlap. The rms values of the processed speech stimuli were normalized with respect to the rms values of the original stimuli.

As shown in Fig. 1, for TESM_E, the entire speech signal was processed with TESM, and then expansion. The rms values of the processed speech stimuli were adjusted to be equal to the rms values of the original stimuli.

B. Stimuli

Speech material consisted of 12/hvd/vowels ([Hillenbrand et al., 1995](#)), 20/aCa/consonants ([Shannon et al., 1999](#)), and 160 IEEE sentences ([Rothausen et al., 1969](#)) spoken by a male speaker. The SSN was constructed by filtering white noise with the talker's long-term speech spectrum derived using a tenth-order linear predictive coding (LPC) analysis. The noise level was varied to generate speech in noise at SNRs of 0, 5, and 10 dB.

Figure 2 shows the temporal waveforms and spectrograms of the original /aKa/ (left panels), /aKa/ after expansion (middle panels), and /aKa/ processed with TESM first, followed by expansion (right panels). It is clear from the

FIG. 2. Temporal waveforms and spectrograms of the consonant/aKa/before expansion (left panel), after expansion (middle panel), and after TESM followed by expansion (right panel).

FIG. 3. Spectra of the vowel in /hid/ in noise at 5 dB SNR before and after processing with E and TESM_E.

temporal waveforms that the short noise burst important for perception of the consonant is attenuated by expansion. The noise burst is recovered to some extent and the onsets of the vowels are also amplified when the speech is processed with TESM first. Spectrograms with E and TESM_E show increases in spectral contrast compared to Original.

Figure 3 shows spectra of the original, expanded, and TESM_E processed vowel /hid/ in SSN at 5 dB SNR. The lighter dotted trace corresponds to the original stimulus, the lighter continuous trace corresponds to the expanded stimulus and the darker trace corresponds to the TESM_E stimulus. Both E and TESM_E increase the spectral contrast (reduce background noise), while preserving the spectral peaks. As TESM is only applied for a small proportion of the signal duration (i.e., during the onset), it is not likely to produce major changes in the spectrum of the entire signal. Thus, E and TESM_E spectra are hardly different.

C. Subjects

Ten CI users between 51 and 82 years old were tested. Subjects included five Nucleus 24 users, two Nucleus 22 users, two Clarion II users, and one Clarion Auria user. Due to limited availability, subject 10 participated in the sentence recognition tests only. Relevant information about all of the subjects is presented in Table I. All subjects were native English speakers. They were paid for participating in the study.

D. Procedure

Subjects were seated in a double-walled, sound attenuating booth during the experiment. In vowel and consonant recognition tests, a graphical user interface consisting of 12 vowels and 20 consonants displayed as pushbuttons was presented on the computer screen. After each stimulus was presented, the subjects were directed to click on the button corresponding to the presented stimulus. Feedback was pro-

vided and the subjects were encouraged to guess if unsure. The noise conditions, including speech in quiet and speech in noise at the different SNRs, were presented randomly and balanced across subjects. For each noise condition, there were four processing conditions (Original, E, TESM and TESM_E), and each stimulus was presented twice for each processing condition. Thus for each noise condition, a total of 96 stimuli for vowel recognition (4 processing conditions $\times$ 2 $\times$ 12 vowels) and 160 stimuli for consonant recognition (4 processing conditions $\times$ 2 $\times$ 20 consonants) were presented randomly. Before testing, the subjects were given $\sim$ 30 min practice to familiarize themselves with the Original stimuli. During practice, a different graphical user interface consisting of vowels and consonants displayed as pushbuttons was presented on the computer screen. The subjects heard the corresponding stimulus when a button was clicked. In the sentence recognition experiment, blocks of 40 sentences were presented for each noise condition, with ten sentences presented for each processing condition. Each sentence consisted of five keywords, producing a total of 50 stimuli for each processing condition. No practice was given.

TABLE I. Information about the subjects.

Subject	Gender	Age (years)	Cause of deafness	Duration of implant use (years)	Clinical strategy
1	F	71	Unknown	4	ACE
2	F	81	Blood clot	5	ACE
3	F	65	Rec. Gene	5	ACE
4	F	58	Unknown	4	ACE
5	F	70	Unknown	7	CIS
6	F	73	Unknown	7	CIS
7	M	51	Trauma	16	SPEAK
8	M	67	Unknown	18	SPEAK
9	M	82	Unknown	6	CIS
10	M	66	Unknown	6	HiRes

The noise conditions, the order of processing and the sentences within each condition were all randomized. Sentences were not repeated. The subjects were asked to type as many words as possible using a computer keyboard after the target sentence was presented. No feedback was provided. The number of correctly identified words was calculated to obtain the final percent correct score. The stimuli were presented via Grason Stadler Clinical Audiometer loudspeakers at 70 dB sound pressure level (SPL), as measured at the subjects' approximate seating positions.

III. RESULTS

Figure 4 shows average percent correct scores and the average changes in percent correct scores as a function of SNR for vowel, consonant and sentence recognition tests, respectively. In general, scores improved with E and TSM_E for all tests in noise. Analysis of variance with subject as the repeated measure was performed on the data from vowel, consonant and sentence recognition tests. The factors were type of processing and SNR (including quiet).

For vowel recognition, there were significant main effects of both factors (type of processing: $F(3,24) = 7.92$, $p < 0.01$, and SNR: $F(3,24) = 17.72$, $p < 0.0005$). There was no significant interaction ($F(9,72) = 1.82$, $p = 0.16$), suggesting that the effect of processing was similar at all SNRs. For pair-wise comparisons of the different types of processing (number of comparisons = 6, using data pooled from all SNRs), the p value was taken as 0.0083, using the Bonferroni correction for multiple comparisons. The improvements in scores with TSM_E relative to both Original ($p = 0.0053$) and TSM ($p = 0.0055$) were signifi-

cant. No significant difference was found between E and Original, E and TSM, E and TSM_E, and TSM and Original.

For consonant recognition, there were significant main effects of type of processing ($F(3,24) = 5.76$, $p < 0.05$) and SNR ($F(3,24) = 51.86$, $p < 0.0005$). There was a significant interaction ($F(3,24) = 6.56$, $p < 0.005$) suggesting that the effect of processing varied with SNR. Comparison of the different processing schemes after Bonferroni correction showed significant improvement with TSM_E over Original ($p = 0.0007$). No significant difference was found between E and Original, E and TSM, and Original and TSM. With E, the mean score in quiet dropped by 13.9 percentage points from the score for Original and the difference was statistically significant using a paired t -test ($p = 0.0023$) after Bonferroni correction (significant $p = 0.0042$ for 12 comparisons). With TSM_E, the mean score improved by 8.6 percentage points over E. However, this difference failed to reach significance ($p = 0.0065$). Also, there was no significant difference between TSM_E and Original scores in quiet.

For sentence recognition, there were significant main effects of processing ($F(3,27) = 10.55$, $p < 0.0005$) and SNR ($F(3,27) = 88.53$, $p < 0.0005$). The interaction was also significant ($p < 0.005$). Multiple comparisons after Bonferroni correction (significant $p = 0.0083$) showed significant improvement with E over Original ($p = 0.0005$) and TSM_E over Original ($p = 0.0075$). The improvement with E over TSM was also significant ($p = 0.0053$). No significant difference was found between TSM_E and TSM ($p = 0.01$), Original and TSM ($p = 0.05$), and E and TSM_E ($p = 0.16$).

FIG. 4. Average percent correct scores (left panels) and average change in percent correct scores (right panel) as a function of SNR for vowel recognition (top panels), consonant recognition (middle panels) and sentence recognition (bottom panels) tasks. (Left panels) Filled circles show average scores with Original stimuli, inverted triangles represent scores with TSM, filled squares show scores with E and diamonds correspond to scores with TSM_E. (Right panels) Using the same symbols as for the processing conditions in the left-hand panel but to represent the difference between the processing condition and the original.

FIG. 5. Sentence recognition scores for individual CI users as a function of SNR in different processing conditions. Filled circles show scores with Original stimuli, inverted triangles show scores with TESM, filled squares show scores with E and diamonds show scores with TESM_E.

To appreciate the effect of the different types of processing on individual CI users, Fig. 5 shows the individual sentence recognition scores. The performance and improvement varied greatly between individual CI users. The performance of CI users fell drastically even in modest levels of background noise. Seven out of ten subjects showed decreases in scores between 24 and 60 percentage points from quiet to 10 dB SNR in the sentence recognition task. Similarly the amount of improvement varied with individual users, ranging from 8 to 64 percentage points with TESM_E and from 14 to 50 percentage points with E.

Finally, to relate different types of processing to enhancement of different cues, sequential information analysis was performed on the consonant confusion matrices pooled across all subjects (Wang and Bilger, 1973). For each processing condition, there were four confusion matrices

corresponding to four SNRs, and each confusion matrix consisted of 360 samples (2 repetitions $\times$ 20 consonants $\times$ 9 subjects). The features voicing, place of articulation, and manner of articulation were divided into two, five, and six categories, respectively.

The information transfer (IT) estimate overestimates the true value when the number of stimulus presentations is small. The overestimation depends upon the number of samples and the number of categories for each feature. For the same number of samples, the overestimate increases with the number of categories (Sagi and Svirsky, 2008). Accordingly, the resultant IT estimate in the present study was within 0.4% of its true value for voicing, 2% for place of articulation, and 2.4% for manner of articulation.

Figure 6 shows the IT estimates at the different SNRs. In quiet (rightmost data points in each panel), E alone

FIG. 6. Information transmission analysis for consonants at different SNRs. The left, middle, and right panels show information transmitted for voicing, place and manner of articulation, respectively. Filled circles correspond to Original stimuli, inverted triangles correspond to TESM, filled squares correspond to E, and diamonds correspond to TESM_E.

(solid squares) produced the lowest information transfer for all features. Compared with the Original condition, TESM alone (inverted open triangles) provided negligible differences in information transmitted for all features. The TESM_E processing (open diamonds) improved information transfer of voicing and manner features over E alone, but information transmission was still less than for both Original and TESM. Overall, none of the three processing schemes had any clear effect on information transfer of the place feature.

IT analysis of the data obtained from speech in noise showed a totally different pattern than for the data in quiet. For all SNRs from 0 to 10 dB, the scores for Original and TESM decreased more with noise than those for E and TESM_E. Comparison of information transmission in quiet and at 10 dB SNR showed that manner of articulation was the most affected by noise, followed by voicing and then place of articulation. More improvement was seen in the transmission of place and manner of articulation than for voicing with E and TESM_E.

IV. DISCUSSION

The present study used expansion (E) to enhance spectral contrast, and TESM to enhance the weak temporal cues associated with certain consonants. The combined spectral and temporal enhancement (TESM_E) scheme showed improvements for all three types of speech materials, and balanced performance between quiet and noise listening conditions.

A. Mechanisms of enhancement and improvement

1. Contrast enhancements with E and TESM

Both E and TESM alter spectrotemporal contrasts, but they produce different patterns of contrast enhancement. E enhances the spectral envelope by expanding the short-duration FFT amplitudes. As the rms levels of the short-duration temporal windows were not preserved, temporal contrast between the windows was expanded as a side effect. On the other hand, TESM enhancement varies from frame to frame only for a dynamically varying signal, and hence depends on the modulation rate in each channel. The gain factor is derived from three consecutive 20 ms segments of the low-pass filtered signal envelope in each channel. The Nyquist rate based on the 20 ms frame width is 25 Hz, and thus envelope modulations rates as high as ~ 25 Hz were expanded by TESM processing.

2. Effect of E and TESM_E on speech reception in quiet

In quiet, the decrease in transmission of voicing information with E and TESM_E may be due to suppression by expansion of the relatively weak harmonics corresponding to voice pitch. This is evident from /g/ being perceived as /k/ with E and /d/ being perceived as /t/ with TESM_E. The main cue to voicing detection in stop consonants is voice onset time, which is shorter for voiced stops than for unvoiced stops. It is possible that perception of voice onset

time, and hence voicing, was more difficult due to suppression of harmonic energy in the consonant-to-vowel transition relative to the preceding vowel resulting from the expansion process. This arose because rms levels were not normalized with respect to unprocessed levels across time frames with the E processing.

The decrease in the transmission of manner information is likely due to changes in the level of weak consonant cues after expansion. This is evident from nasals being perceived as fricatives after expansion with E (e.g., /m/ was perceived as /v/). Nasals are characterized by a large concentration of acoustic power in the low-frequency region, weak but distinct second and third formant frequencies, and a spectral zero. The voiced fricative /v/ has acoustic energy in the low-frequency region and frication spread across the high-frequency region (Levitt, 1978). Often, the voicing in /v/ is more powerful than the frication and its spectrogram appears similar to that of nasals with suppressed second and third formant frequencies. Improvement in the transmission of manner with TESM_E over E in quiet was mainly seen as improvement in the perception of fricatives, stops, and nasals. This indicates that the low-intensity short-duration cues were indeed enhanced by TESM, and for TESM_E, this compensated in part for distortions in some vowel-to-consonant intensity ratios introduced by E processing.

Overall, the decrease in intelligibility of the consonants in quiet with E can be attributed to the reduced consonant–vowel intensity ratio. Although not significant, enhancing the consonant–vowel intensity ratio (by increasing the consonant intensity with TESM) in TESM_E, improved the perception of consonants with TESM_E compared to E by 8.6 percentage points. This trends toward the conclusion from earlier studies that increasing the consonant–vowel intensity ratio can improve the perception of consonants (Gordon-Salant, 1986; Kennedy *et al.*, 1998; Montgomery and Edge, 1988).

3. Effect of E and TESM_E on the reception of speech in noise

Decrease in the transmission of features with Original and TESM produced by adding noise was likely due to decrease in both temporal and spectral contrast. Improvement in the transmission of manner with TESM_E in noise was likely due to enhanced coding of consonant cues. Figure 7 shows temporal waveforms and spectrograms of the Original, TESM processed, E processed, and TESM_E processed consonant /aKa/ in SSN at 5 dB SNR. TESM_E not only appears to suppress the background noise, but also enhances the short duration cue important for the perception of the consonant. Improvement in the place of articulation probably results from improvement in the spectral contrast with both E and TESM_E. Vowel perception in noise also improved with TESM_E, which was likely due to increase in spectral contrast produced by expansion (Fig. 3).

4. Benefits of E and TESM_E in sentence recognition

More improvement with E and TESM_E was seen for sentences in noise than for vowels and consonants in noise. It is likely that this was due to suppression of noise in the

FIG. 7. Temporal waveforms and spectrograms of the consonant in/aKa/in SSN at 5 dB SNR: Original (leftmost panel), TESM processed (second panel), E processed (third panel), and TESM_E processed (rightmost panel).

intervening low-amplitude regions between words in the sentences due to enhancement of temporal contrast between frames. Improvement with TESM_E is also likely due to the enhancement of weak consonants, particularly the initial consonants in words. According to the lexical-access model proposed by Stevens (2002), consonants serve as important landmarks in a speech stream. The word-initial consonants are critically important for defining the word onsets, and contain several landmark features. Absence of these landmark features in noise can disrupt perception of the syllable structure, and blur the word onset. Enhanced access to the word-initial consonants improves identification of the word boundaries, and consequently improves word identification in noise (Li and Loizou, 2008).

B. Comparison with previous enhancement studies

With Bunnell's (1990) contrast enhancement technique, HI listeners with sloping loss showed small improvements in identifying stop consonants presented in quiet. The contrasts were altered primarily in the mid-frequency region, leaving the low- and high-frequency regions unaffected. In the present study, spectral expansion was applied over the entire frequency range, and outcomes were substantially different to those observed by Bunnell, perhaps in part due to subject differences between studies (HI listeners with sloping loss in Bunnell's study and CI users in the present study). Baer *et al.* (1993) showed that their enhancement technique produced small (1.76%), but statistically significant improvement in sentence verification in HI users with moderate degree of bilateral sensorineural hearing loss. The enhancement technique involved obtaining the auditory excitation pattern by transforming the magnitude spectrum and then convolving the auditory excitation pattern with a "difference

of Gaussians" function. The present study showed larger improvements in a sentence identification test, and differed substantially from the study by Baer *et al.* in terms of processing, speech material (Adaptive Sentence Lists) in Baer *et al.* (1993), and vowels, consonants, and IEEE sentences in the present study] and subjects [HI listeners with moderate degree of bilateral sensorineural hearing loss in Baer *et al.*, (1993), and CI users in the present study].

Lyzenga *et al.* (2002) reported no improvement in speech reception threshold (SRT) scores with spectral expansion. However, the present study showed significant improvement in sentence intelligibility using spectral expansion. The different outcomes can possibly be attributed to the implementation differences between the two studies, and/or to subjects that were NH listeners listening to simulations of reduced frequency selectivity in Lyzenga *et al.* (2002) and CI users in this study.

Franck *et al.* (1999) investigated the effect of spectral expansion, spectral expansion combined with a multichannel phonemic compressor, and spectral expansion combined with a single-channel phonemic compressor. Speech intelligibility was tested both in quiet and in noise using monosyllabic (CVC) words. They found that, in quiet, the consonant scores with unprocessed stimuli tended to be higher than for the spectrally expanded stimuli, which is consistent with the outcomes of the present study.

Spectral enhancement has previously been shown to improve speech intelligibility for CI users listening in noise (Bhattacharya and Zeng, 2007; Oxenham *et al.*, 2007). Bhattacharya and Zeng (2007) evaluated the effect of a spectral enhancement technique, called Companding, in vowel, consonant, and sentence recognition tasks. For CI users, the maximum average improvement (based on average improvements for the different SNRs) in vowel, consonant and sentence

recognition was 21, 12 and 18 percentage points, respectively. The present study showed similar improvements in sentence recognition and smaller improvements in vowel and consonant recognition. An advantage of E and TESM_E over Compadding is lower computational complexity.

Although the study of Vandali (2001) showed significant improvements with TESM over Original in consonant and vowel scores for a consonant–vowel nucleus–consonant (CNC) word test in quiet, the present study failed to show any significant improvement in the perception of consonants with TESM. A possible reason for this discrepancy is the lower stimulation rate (400 Hz), and lower front-end sensitivity of the processor used by Vandali (2001) compared to the current clinical settings used here. This agrees with the study of Holden *et al.* (2005) comparing ACE (Advanced Combination Encoder) and TESM, which showed lower benefits with TESM over ACE than those reported by Vandali comparing spectral maxima sound processor and TESM. The discrepancy was attributed to the higher stimulation rates and higher front-end sensitivity of the processors used by Holden *et al.* (2005) compared to those used by Vandali. The higher sensitivity provided access to lower level consonant information, and hence reduced potential benefits to consonant perception provided by TESM. Benefits of using TESM were observed only for speech presented at soft levels, below 60 dB SPL (Holden *et al.*, 2005). In the current study, stimuli were presented at 70 dB SPL. Thus, consonant information was presented at sufficiently audible levels, which explains the lack of benefits observed using TESM. Other differences include different degrees of gain applied to the short-duration cues ($k=2.5$ in the current study, and $k=2$ in Vandali's study). Also, the consonant and vowel perception scores were derived from monosyllabic CNC word tests in Vandali's study, which are much harder than the closed set vowel and consonant tests used in this study.

C. Applications

The present study showed that preprocessing speech with E or TESM_E can improve the ability of CI users to understand speech in steady speech-shaped noise. However, these strategies remain to be tested in fluctuating background noise such as multitalker babble. It is unlikely that increasing the spectral contrast will improve the SNR in fluctuating noise conditions because of the dynamic nature of both the signal and the noise.

The simplicity of implementation makes E or TESM_E ideal for incorporation as a front-end to CI processing. The effectiveness of these strategies as a front-end to hearing aid processing remains to be explored. Multichannel compression in hearing aids reduces the spectral and temporal contrast (Plomp, 1988), which in turn affects speech perception in noise. E and TESM_E have the potential to overcome this problem. However, the E and TESM_E parameters, viz., the FFT lengths, gain factor in TESM and the degree of expansion in E, need to be carefully coordinated with the hearing aid processing parameters, viz., number of channels, amount of compression, and the attack and release times.

In the present study, the rms level of the entire processed stimulus was set equal to the rms level of the original

stimulus, rather than equalizing on a frame-by-frame basis. However, in a real-time implementation of the system, it is not possible to do this. Instead rms level could be maintained by applying a fixed amount of gain to the processed signal for all frames. The amount of gain could be chosen so as to preserve rms levels of long-term average speech presented at a comfortable listening level (e.g., 60–65 dB SPL). Alternatively, the gain could be slowly adapted based on a long-term average of the incoming speech level.

Further work is needed to reduce the delay in processing introduced by TESM_E. In the current study, for E, the delay associated with a sampling rate of 20 kHz and FFT length of 512 is 25.6 ms. This, added to the 30 ms delay introduced by TESM in TESM_E, amounts to 55.6 ms delay, which would be noticeable in a real-time implementation (Vandali, 2001). The delay could be reduced by decreasing the FFT length and/or decreasing the sampling rate. However, decreasing the FFT length alone would reduce the frequency resolution. Also, the sampling rate cannot be decreased drastically because doing so would reduce the audio bandwidth and degrade the speech quality. A better solution for reducing the delay is to combine TESM and E in the same process, thereby fixing the delay to the longer of the two individual processing delays. The TESM delay could be reduced slightly by determining the slowly varying envelopes via averaging over 18 ms durations, thereby fixing the delay of both TESM and E to 25.6 ms. Stone and Moore (2005) found that hearing-aid processing delays up to 32 ms were acceptable to HI listeners with severe low-frequency hearing loss. Vandali (2001) found that for CI users, most subjects became accustomed to delays of 30 ms in TESM processing within a short period of time although a few never quite became accustomed to it (Holden *et al.*, 2005).

An alternative to adding TESM before E would be to use a SNR detector to reduce the exponent used in E when high SNRs are measured. Apart from their use in assistive listening devices, E or TESM_E may also be used in applications like automatic speech recognition systems that are very susceptible to background noise.

V. CONCLUSIONS

Three spectral and temporal contrast enhancement strategies were evaluated for cochlear-implant speech recognition in quiet and in noise. The three strategies included spectral expansion (E) to enhance spectral contrast, TESM to enhance temporal transients, and the combined TESM_E. The following conclusions can be drawn:

- (1) E processing improved cochlear-implant speech recognition in noise but degraded it in quiet, particularly for consonant recognition.
- (2) The deleterious effect of E on consonant recognition could be partially counteracted by adding TESM before E
- (3) The combined TESM_E processing showed improvement for all three types of speech materials, and balanced performance between quiet and noise listening conditions.
- (4) The amount of improvement varied greatly across individual CI users, suggesting the need for customized processing to optimize performance.

ACKNOWLEDGMENTS

The authors would like to thank all the subjects who participated in the experiments. They also thank the anonymous reviewers for their helpful comments. This work was supported by NIH (2R01 DC008858 and 1P30 DC008369).

- Baer, T., Moore, B. C. J., and Gatehouse, S. (1993). "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times," *J. Rehabil. Res. Dev.* **30**, 49–72.
- Bhattacharya, A., and Zeng, F. G. (2007). "Companding to improve cochlear-implant speech recognition in speech-shaped noise," *J. Acoust. Soc. Am.* **122**, 1079–1089.
- Bunnell, H. T. (1990). "On enhancement of spectral contrast in speech for hearing-impaired listeners," *J. Acoust. Soc. Am.* **88**, 2546–2556.
- Busby, P. A., Tong, Y. C., and Clark, G. M. (1993). "Electrode position, repetition rate, and speech perception by early- and late-deafened cochlear implant patients," *J. Acoust. Soc. Am.* **93**, 1058–1067.
- Chatterjee, M., and Shannon, R. V. (1998). "Forward masked excitation patterns in multielectrode electrical stimulation," *J. Acoust. Soc. Am.* **103**, 2565–2572.
- Clarkson, P. M., and Bahgat, S. F. (1991). "Envelope expansion methods for speech enhancement," *J. Acoust. Soc. Am.* **89**, 1378–1382.
- Cohen, L., Xu, J., Xu, S. A., and Clark, G. M. (1996). "Improved and simplified methods for specifying positions of the electrode bands of a cochlear implant array," *Am. J. Otol.* **17**, 859–865.
- Cohen, L. T. (2009). "Practical model description of peripheral neural excitation in cochlear implant recipients: 2. Spread of the effective stimulation field (ESF), from ECAP and FEA," *Hear. Res.* **247**, 100–111.
- Cohen, L. T., Richardson, L. M., Saunders, E., and Cowan, R. S. C. (2003). "Spatial spread of neural excitation in cochlear implant recipients: Comparison of improved ECAP method and psychophysical forward masking," *Hear. Res.* **179**, 72–87.
- Dorman, M. F., Loizou, P. C., Fitzke, J., and Tu, Z. (1998). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels," *J. Acoust. Soc. Am.* **104**, 3583–3585.
- Franck, B. A. M., van Kreveld-Bos, C. S. G. M., Dreschler, W. A., and Verschuure, H. (1999). "Evaluation of spectral enhancement in hearing aids, combined with phonemic compression," *J. Acoust. Soc. Am.* **106**, 1452–1464.
- Gordon-Salant, S. (1986). "Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing," *J. Acoust. Soc. Am.* **80**, 1599–1607.
- 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.
- Hochberg, I., Boothroyd, A., Weiss, M., and Hellman, S. (1992). "Effects of noise and noise suppression on speech perception by cochlear implant users," *Ear Hear.* **13**, 263–271.
- Holden, L. K., Vandali, A. E., Skinner, M. W., Fourakis, M. S., and Holden, T. A. (2005). "Speech recognition with the advanced combination encoder and transient emphasis spectral maxima strategies in nucleus 24 recipients," *J. Speech Lang. Hear. Res.* **48**, 681–701.
- Kennedy, E., Levitt, H., Neuman, A. C., and Weiss, M. (1998). "Consonant-vowel intensity ratios for maximizing consonant recognition by hearing-impaired listeners," *J. Acoust. Soc. Am.* **103**, 1098–1114.
- Levitt, H. (1978). "Acoustics of speech production," in *Auditory Management of Hearing-Impaired Children*, edited by M. Ross and T. Giolas (University Park Press, Baltimore, MD), pp. 45–115.
- Li, N., and Loizou, P. C. (2008). "The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise," *J. Acoust. Soc. Am.* **124**, 3947–3958.
- Loizou, P. C., Dorman, M., and Fitzke, J. (2000). "The effect of reduced dynamic range on speech understanding: Implications for patients with cochlear implants," *Ear Hear.* **21**, 25–31.
- Loizou, P. C., and Poroy, O. (2001). "Minimum spectral contrast needed for vowel identification by normal hearing and cochlear implant listeners," *J. Acoust. Soc. Am.* **110**, 1619–1627.
- Lyzenga, J., Festen, J. M., and Houtgast, T. (2002). "A speech enhancement scheme incorporating spectral expansion evaluated with simulated loss of frequency selectivity," *J. Acoust. Soc. Am.* **112**, 1145–1157.
- Montgomery, A. A., and Edge, R. A. (1988). "Evaluation of two speech enhancement techniques to improve intelligibility for hearing-impaired adults," *J. Speech Hear. Res.* **31**, 386–393.
- Nelson, D. A., Schmitz, J. L., Donaldson, G. S., Viemeister, N. F., and Javel, E. (1996). "Intensity discrimination as a function of stimulus level with electric stimulation," *J. Acoust. Soc. Am.* **100**, 2393–2414.
- Nelson, P. B., Jin, S.-H., Carney, A. E., and Nelson, D. A. (2003). "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," *J. Acoust. Soc. Am.* **113**, 961–968.
- Oxenham, A. J., Simonson, A. M., Turicchia, L., and Sarpeshkar, R. (2007). "Evaluation of companding-based spectral enhancement using simulated cochlear-implant processing," *J. Acoust. Soc. Am.* **121**, 1709–1716.
- Plomp, R. (1988). "The negative effect of amplitude compression in multi-channel hearing aids in the light of the modulation-transfer function," *J. Acoust. Soc. Am.* **83**, 2322–2327.
- Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H., Urbanek, G. E., and Weinstock, M. (1969). "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.* **17**, 227–246.
- Sagi, E., and Svirsky, M. A. (2008). "Information transfer analysis: A first look at estimation bias," *J. Acoust. Soc. Am.* **123**, 2848–2857.
- Shannon, R. V., Jansvold, A., Padilla, M., Robert, M. E., and Wang, X. (1999). "Consonant recordings for speech testing," *J. Acoust. Soc. Am.* **106**, L71–L74.
- Stevens, K. N. (2002). "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Am.* **111**, 1872–1891.
- Stickney, G. S., Zeng, F.-G., Litovsky, R., and Assmann, P. (2004). "Cochlear implant speech recognition with speech maskers," *J. Acoust. Soc. Am.* **116**, 1081–1091.
- Stone, M. A., and Moore, B. C. J. (2005). "Tolerable hearing-aid delays: IV. Effects on subjective disturbance during speech production by hearing-impaired subjects," *Ear Hear.* **26**, 225–235.
- Vandali, A. E. (2001). "Emphasis of short-duration acoustic speech cues for cochlear implant users," *J. Acoust. Soc. Am.* **109**, 2049–2061.
- Wang, M. D., and Bilger, R. C. (1973). "Consonant confusions in noise: A study of perceptual features," *J. Acoust. Soc. Am.* **54**, 1248–1266.
- Zeng, F. G., and Galvin, J. J. (1999). "Amplitude mapping and phoneme recognition in cochlear implant listeners," *Ear Hear.* **20**, 60–74.
- Zeng, F. G., Nie, K., Stickney, G. S., Kong, Y.-Y., Vongphoe, M., Bhargave, A., Wei, C., and Cao, K. (2005). "Speech recognition with amplitude and frequency modulations," *Proc. Natl. Acad. Sci. U.S.A.* **102**, 2293–2298.