

Effects of introducing low-frequency harmonics in the perception of vocoded telephone speech^{a)}

Yi Hu^{b)}

Department of Electrical Engineering and Computer Science, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53201

Philipos C. Loizou

Department of Electrical Engineering, The University of Texas-Dallas, Richardson, Texas 75080

(Received 4 November 2009; revised 2 June 2010; accepted 15 June 2010)

Several studies have demonstrated that telephone use presents a challenge for most cochlear implant (CI) users, and this is attributed mainly to the narrow bandwidth (300–3400 Hz) introduced by the telephone network. The present study focuses on answering the question whether telephone speech recognition in noise can be improved by introducing, prior to vocoder processing, low-frequency harmonic information encompassing the missing (due to the telephone network) information residing in the 0–300 Hz band. Experiment 1 regenerates the main harmonics and adjacent partials within the 0–600 Hz range in corrupted (by steady noise) telephone speech which has been vocoded to simulate electric-acoustic stimulation (EAS). Results indicated that introducing the main harmonics alone did not produce any benefits in intelligibility. Substantial benefit (20%) was observed, however, when both main harmonics and adjacent partials were regenerated in the acoustic portion of EAS-vocoded telephone speech. A similar benefit was noted in Experiment 2 when low-frequency harmonic information was introduced prior to processing noise-corrupted telephone speech using an eight-channel vocoder. The gain in telephone speech intelligibility in noise obtained when low-frequency harmonic information was introduced can be attributed to the listeners having more reliable access to a combination of F0, glimpsing and lexical segmentation cues.

© 2010 Acoustical Society of America. [DOI: 10.1121/1.3463803]

PACS number(s): 43.66.Ts, 43.71.Ky [MW]

Pages: 1280–1289

I. INTRODUCTION

The public telephone network passes only spectral information within the range of 300 Hz to 3400 Hz. This limited telephone bandwidth (300–3400 Hz) severely limits the ability of normal-hearing and hearing-impaired listeners to identify high-frequency consonants (e.g., /s/) or discriminate between certain consonant pairs such as /s/ and /f/, or /p/ and /t/. According to Kepler *et al.* (1992), this reduced bandwidth is one of the three main factors contributing to the telephone communication difficulty experienced by the hearing-impaired population. Several studies have been conducted to evaluate telephone usage by hearing-impaired people wearing hearing aids or cochlear implants (Cohen *et al.*, 1989; Kepler *et al.*, 1992; Cray *et al.*, 2004), and in general, hearing-impaired people showed strong interest in wanting to use the telephone. Kepler *et al.* (1992) reported that 55% of the respondents who wore hearing aids used their aids while talking on the phone, and 75% indicated an interest in improvements for hearing-impaired people's telephone usage. Cray *et al.* (2004) conducted an investigation of telephone use among cochlear implant recipients, and found that 70% of the respondents who wore cochlear implants were tele-

phone users, 30% of which communicated via a cellular phone for personal use. These surveys indicated a need to improve the ability of hearing-impaired people to communicate via the telephone.

A number of previous studies (Ito *et al.*, 1999; Milchard and Cullington, 2004; Fu and Galvin, 2006; Horng *et al.*, 2007) that assessed the CI listeners' capacity to use telephone have demonstrated that although CI listeners can use the telephone to some degree, their recognition of telephone speech is significantly worse compared with their recognition of wideband speech. The most likely explanation is that telephone speech does not convey information below 300 Hz and above 3400 Hz, which is problematic as speech information below 300 Hz contains F0 cues, and information above 3400 Hz is useful in identifying certain high-frequency consonants (e.g., /s/). Another important conclusion drawn by Fu and his colleagues (Fu and Galvin, 2006; Horng *et al.*, 2007) is that there exists significant inter-subject variability due to the reduced speech bandwidth, as CI users show different capabilities in making use of high-frequency speech cues.

Only a few studies have attempted to develop methods to improve the recognition of telephone-processed speech by hearing impaired listeners. Ito *et al.* (1999) reported that using telephone adapters could improve telephone speech intelligibility by CI users, although these devices are not readily available. In Terry *et al.* (1992), two digital signal

^{a)} Portions of this work were presented at the 33rd annual Midwinter Research Meeting of the Association for Research in Otolaryngology.

^{b)} Author to whom correspondence should be addressed. Electronic mail: huy@uwm.edu

processing techniques designed to compensate for high-frequency hearing loss were shown to significantly improve telephone speech intelligibility for the hearing impaired. A recent study by [Liu et al. \(2009\)](#) utilized bandwidth extension techniques to partly restore high-frequency information for telephone speech. Their results showed a modest but significant improvement in telephone speech recognition by 7 CI users with the proposed method. Both methods proposed by [Terry et al. \(1992\)](#) and [Liu et al. \(2009\)](#) were based on algorithms that improved the representation of the high-frequency information.

A different approach is taken in the present study. Rather than focusing on introducing information in the high-frequency region (>3.4 kHz) of the spectrum as done in the aforementioned studies, we instead focus on introducing low-frequency information (<600 Hz), which encompasses the missing (due to the telephone network) information contained in the 0–300 Hz band. The 600-Hz cutoff was chosen because many CI users with residual hearing are now fitted with both hearing aids and cochlear implants ([Turner et al., 2004](#); [Kong et al., 2005](#); [Gantz et al., 2005, 2006](#); [Dorman et al., 2008](#)). A number of studies investigated the improved speech intelligibility when low-frequency information is combined with the electrical information presented via the CI. Large benefits were noted, particularly in noisy backgrounds, and most of the benefits were attributed to either a better representation of F0 cues and/or better access to glimpsing cues ([Kong and Carlyon, 2007](#); [Li and Loizou, 2008b](#); [Brown and Bacon, 2009b](#)). Given the contribution and benefit of these cues to speech recognition in noise, the present study examines whether introducing low-frequency information enhances the perception of CI-vocoded or EAS-vocoded telephone speech. We thus considered introducing low-frequency information in the 0–300 Hz region by applying harmonics regeneration techniques in addition to the otherwise available (but moderately distorted) 300–600 Hz region. The harmonics-regeneration concept was introduced in [Hu \(2010\)](#) and [Hu and Loizou \(2010\)](#) and applied to the situation in which the noisy signal was first processed by a noise-suppression algorithm and subsequently EAS vocoded ([Hu, 2010](#)) or CI vocoded ([Hu and Loizou, 2010](#)). Hence, the harmonics-regeneration approach was applied in our prior studies in the context of noise reduction (not used in the present study). Unlike the two previous studies which focused on noise reduction, the present study focuses on the intelligibility of speech processed by the telephone network, which effectively eliminates low-frequency (<300 Hz) and high frequency (>3.4 kHz) information. This makes the telephone-listening scenario an ideal scenario for the harmonics-regeneration approach, as it assesses the potential benefit of introducing back the missing harmonics in the low frequencies.

Low-frequency harmonic information is introduced for several reasons. First, F0 has been shown previously to contribute significantly to the benefit associated with EAS ([Qin and Oxenham, 2006](#); [Zhang et al., 2010](#)). Second, from a practical point of view, it would be easier to use signal processing techniques to regenerate information in the low-frequency region (0–600 Hz) rather than regenerate informa-

tion in the high-frequency region (>3.4 kHz) which spans as much as an octave above 3.4 kHz. This is so because in noisy conditions, the low-frequency information (e.g., F0/F1) is shielded to some extent ([Parikh and Loizou, 2005](#)), thus making the detection and estimation of low-frequency components (e.g., harmonic components) much easier than the detection/estimation of high-frequency components. Low frequency (<600 Hz) information contained in voiced segments (e.g., vowels) is mostly composed of harmonics, while high frequency (>3.4 kHz) information contained in unvoiced segments (e.g., fricatives) in speech is mostly composed of noise-like components, which are difficult to recover. The underlying hypothesis in the present study is that by introducing low-frequency information in noisy conditions (particularly in steady background noise which lacks temporal gaps), we can provide a better representation of F0 information, F1 information and/or better access to glimpsing cues ([Kong and Carlyon, 2007](#); [Li and Loizou, 2008b](#); [Brown and Bacon, 2009b](#)). It is not clear whether an improved representation of low-frequency information will benefit perception of telephone speech in noise. This question is investigated in the present study using vocoded telephone speech. Two experiments were conducted to assess the benefits of introducing low-frequency information to EAS-vocoded telephone speech (Exp. 1) as well as CI-vocoded telephone speech (Exp. 2).

II. EXPERIMENT 1: EFFECTS OF INTRODUCING LOW-FREQUENCY HARMONIC INFORMATION IN EAS-VOCODED TELEPHONE SPEECH

A. Methods

1. Subjects and stimuli

Seven normal-hearing native speakers of American English participated in this experiment. All subjects were paid for their participation, and all of them were undergraduate and graduate students at the University of Texas-Dallas. The target speech materials consisted of sentences from the IEEE database ([IEEE, 1969](#)) and were obtained from [Loizou \(2007\)](#). The IEEE corpus contains 72 lists of ten phonetically balanced sentences produced by a male speaker and recorded in a double-walled sound-attenuation booth at a sampling rate of 25 kHz. The masker was steady speech-shaped noise and had the same long-term spectrum as the sentences in the IEEE corpus.

To simulate the receiving frequency characteristics of telephone handsets, the speech and the speech-shaped noise signals were filtered by the modified Intermediate Reference System (IRS) filters used in ITU-T [P.862 \(2000\)](#). The frequency response of the filter is shown in [Fig. 1](#). The filtered speech-shaped noise was added to the filtered speech signal at -1 , 1 , 3 and 5 dB, and the degraded speech signals were downsampled to 8 kHz. The main reason we studied telephone speech perception in noise is that telephone speech can be degraded by noise, and there are several possible scenarios: first, noise can be introduced at the transmitter end (e.g., when the other talker is using a cellular phone); sec-

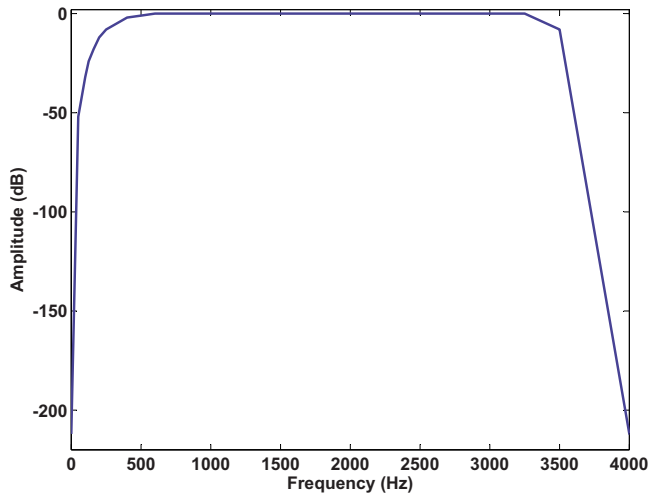


FIG. 1. (Color online) Frequency response of IRS filter simulating frequency characteristics of telephone handsets.

ond, many CI and EAS users use the loudspeaker on the phone, and in this scenario ambient noise will degrade the speech information.

2. Signal processing

The experiments were designed to evaluate the benefits of introducing low-frequency harmonics in the acoustic portion of simulated EAS processing. Figure 2 shows the block diagram of the overall system. A total of three processing conditions were used for this purpose. The first processing condition was designed to simulate combined electric and acoustic stimulation, henceforth referred to as EAS processing. As the first step, a pre-emphasis filter with 2000 Hz cutoff frequency and 3 dB/octave rolloff was applied to the signal. For the acoustic stimulation, the speech signal was lowpass filtered at 600 Hz using a sixth-order Butterworth filter. For the electric simulation, a five-channel noise-excited vocoder was used (Shannon *et al.*, 1995). The speech signal was bandpassed into five frequency bands using sixth-order Butterworth filters allocated according to the Equivalent Rectangular Bandwidth (ERB) filter spacing (Glasberg and Moore, 1990) (see Table I). The envelopes of the bandpassed signals were obtained by full-wave rectification, followed by low-pass filtering using a second-order Butterworth filter with a 400 Hz cutoff frequency. Independent white-noise carriers were used to modulate the temporal envelopes extracted in each band. Following the modulation, the same bandpass filters (as in Table I) were used to filter the amplitude-modulated envelopes. The narrow-band modulated signals were finally summed across all bands and combined with the lowpass filtered speech signal to generate the EAS stimuli.

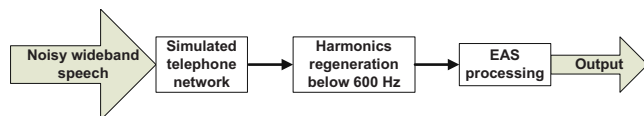


FIG. 2. (Color online) Block diagram of the overall system for simulated EAS processing.

TABLE I. Cutoff frequencies of filters used in simulated EAS processing.

Channel	Lowpass filtering below 600 Hz	
	Low (Hz)	High (Hz)
1	549	830
2	830	1212
3	1212	1731
4	1731	2438
5	2438	3400

The level of the synthesized speech signal was scaled to have the same root mean square (RMS) value as the original speech signal.

The other two conditions were designed to evaluate the benefits of introducing non-distorted low-frequency harmonics in the acoustic portion of the EAS stimuli. In order to establish an upper bound and evaluate the potential of the harmonics regeneration stage, we assumed an ideal operating environment. That is, we estimated the F0 value from the wideband clean speech signal and regenerated the signal's harmonics with prior knowledge of the wideband clean speech spectrum. Figure 3 shows the block diagram for the two conditions. The speech signals were segmented into 50% overlapping frames using a 20-ms Hanning window. The fast Fourier transform (FFT) was applied to each 20-ms frame. An F0-detection algorithm based on the autocorrelation function (Kondoz, 1999) was used in the present study. If the F0 value of a frame was found to be greater than 75 Hz and smaller than 500 Hz, the frame was detected to be voiced; otherwise the frame was declared to be unvoiced and passed through without further processing. In the voiced frames, the

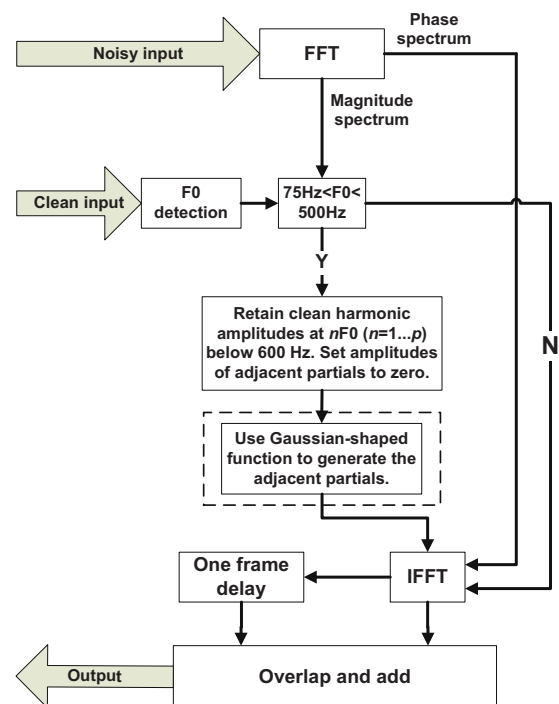


FIG. 3. (Color online) Block diagram of the two processing conditions used for generating the HAR and HAR+PAR stimuli.

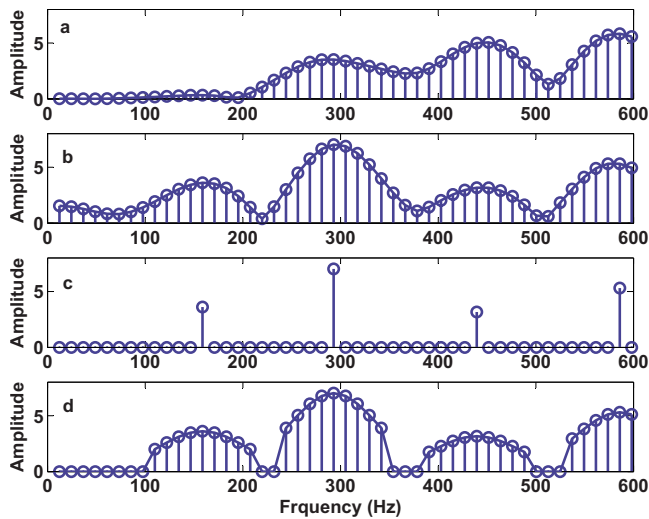


FIG. 4. (Color online) (a) FFT magnitude spectrum of a voiced segment extracted from telephone-processed speech. (b) FFT magnitude spectrum of a voiced segment extracted from wideband speech. (c) Example HAR stimulus obtained by sampling the spectrum shown in (b) at multiples of F_0 . (d) Example HAR+PAR stimulus obtained by retaining the main harmonics and generating the adjacent partials using the Gaussian-shaped function centered around each harmonic.

FFT magnitude spectrum was sampled at integer multiples of F_0 to extract the harmonic amplitudes. The number of regenerated harmonics was calculated by $p = \lfloor CF/F_0 \rfloor$, where CF is the cutoff frequency below which harmonics are included, and the symbol $\lfloor \cdot \rfloor$ indicates the flooring operation. A cutoff frequency of 600 Hz was used for two reasons: first, the telephone-network filter starts sloping around 500 Hz; second, regenerating harmonics below 600 Hz can take full advantage of the acoustic bandwidth available to EAS users (Dorman *et al.*, 2008, Fig. 1). To compensate for the possible inaccuracy of the F_0 detector, harmonics were regenerated by extracting the local peaks in a 30-Hz range around nF_0 , where $n = 1, \dots, p$.

The magnitude spectra of voiced phonetic segments (e.g., vowels) possess a harmonic structure consisting of both harmonics and adjacent partials (see example in Fig. 4). The adjacent partials, falling between the main harmonic components (which occur primarily at multiples of F_0), are also present in voiced magnitude spectra (Fig. 4, panel b). To assess the importance of preserving the harmonic structure of voiced segments, two conditions were created. In the first condition, only the main harmonic amplitudes were regenerated, and the amplitudes of the adjacent partials were set to zero (Fig. 4, panel c). In the second condition, the amplitudes of the main harmonics along with the adjacent partials were regenerated. We denote the first condition as HAR, and the second condition in which both the main harmonics and the adjacent partials are included as HAR+PAR. For comparative purposes, the control EAS condition is also included in the listening tests. Note that the input stimuli in the control EAS condition is telephone-bandwidth speech. Following the extraction of the harmonic amplitudes (along with the generated partials in the HAR+PAR condition), an inverse FFT is applied to the modified magnitude spectrum using the noisy-speech phase spectrum, and the overlap-and-add procedure is finally applied.

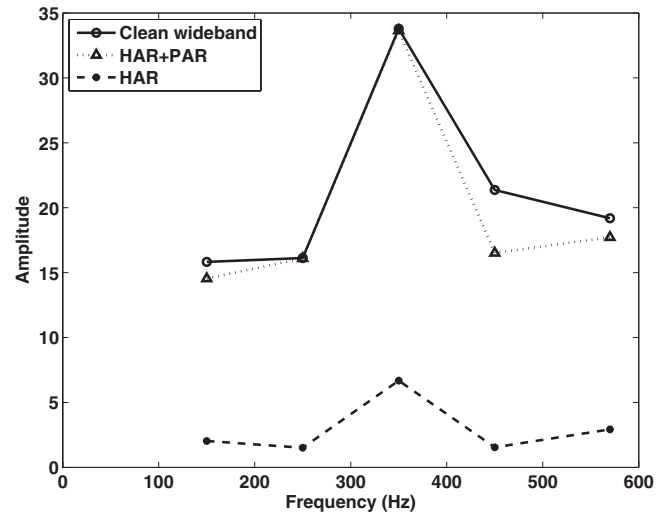


FIG. 5. Excitation spectra of the HAR, HAR+PAR and wideband clean stimuli shown in Fig. 4.

In the HAR+PAR condition, the partials were not extracted from the clean spectra. Instead, a simple approach was taken to generate the neighboring partials. This was done by multiplying the main harmonic amplitudes by a Gaussian-shaped function, and sampling the Gaussian function at 16 discrete frequencies (spanning a total bandwidth of 100 Hz) to the left and right of the main harmonics. The Gaussian-window bandwidth (100 Hz) was chosen to accommodate the F_0 of the male speaker. The Gaussian function was derived heuristically by inspecting the magnitude spectra of several frames of voiced segments. More complex algorithms could alternatively be used to generate the Gaussian function, however we chose the Gaussian function for its simplicity and practical implications. In a realistic implementation, the neighboring partials do not need to be estimated from the noisy signal, only the main harmonics need to be estimated.

Figure 4 shows example HAR and HAR+PAR stimuli extracted from a voiced speech segment (only the acoustic portion is displayed). The HAR stimuli consist primarily of the main harmonics (the adjacent partials are set to zero). In contrast, the HAR+PAR stimuli consist of the main harmonics (occurring at multiples of F_0) and adjacent partials which are generated using a Gaussian-shaped function. Comparing panels (b) and (d), it is clear that the HAR+PAR stimuli approximate well the clean low-pass acoustic information in the 0–600 Hz range. This was also confirmed by examining the excitation spectra of the wideband, HAR and HAR+PAR stimuli. The excitation spectra, computed as per Moore and Glasberg (1983), of the HAR+PAR stimuli approximate well those of the wideband stimulus (see Fig. 5). Note that the second prominent harmonic is captured well in the excitation spectra of the HAR+PAR stimulus. While the second harmonic is also evident in the excitation spectra of the HAR stimulus, it is not as distinct due to the reduced spectral contrast.

3. Procedure

The listening tests were conducted using a PC connected to a Tucker-Davis system 3. Stimuli were played monaurally

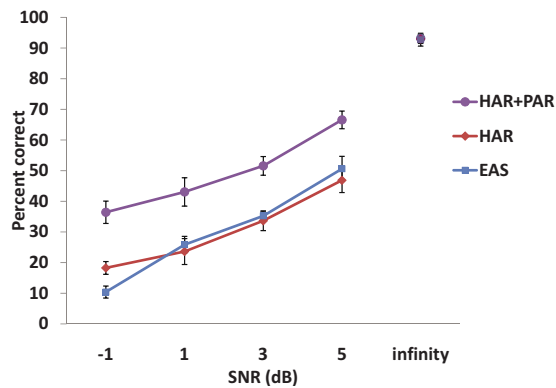


FIG. 6. (Color online) Mean percent correct scores as a function of SNR level. The error bars denote ± 1 standard error of the mean.

to the subjects through Sennheiser HD 250 Linear II circumaural headphones at a comfortable listening level. The subjects were seated in a double-walled sound-attenuation booth (Acoustic Systems, Inc.). To familiarize each subject with the stimuli, a training session was administered prior to the formal testing, and each subject listened to EAS-processed speech stimuli during the training session, which typically lasted about 15 to 20 min. During the testing session, the subjects were instructed to type the words they heard in a Matlab Graphic User Interface (GUI) program. In total, there were 12 testing conditions ($=4$ SNR levels $\times 3$ processing methods). For each condition, two lists of sentences were used, and none of the lists were repeated across the testing conditions. The conditions were presented in random order for each subject. The subjects were allowed to take breaks at their leisure and no feedback was provided after each testing condition.

B. Results

The mean percent correct scores for all conditions in EAS processing are shown in Fig. 6. Performance was measured in terms of percent of words identified correctly (all words were scored). The scores were first converted to rational arcsine units (RAU) using the rationalized arcsine transform proposed by Studebaker (1985). To examine the effect of SNR level and type of harmonic structure (main harmonics only vs. main harmonics plus adjacent partials, i.e., HAR vs. HAR+PAR) regenerated, we subjected the RAU-converted scores to statistical analysis using the converted score as the dependent variable, and the SNR level and type of harmonic structure as the two within-subjects factors. Analysis of variance (ANOVA) with repeated measures indicated significant effects of SNR level [$F(3, 18)=39.95, p < 0.0005$] and type of harmonic structure [$F(2, 12)=81.41, p < 0.0005$]. There were significant between-factor interactions between SNR level and type of harmonic structure [$F(6, 36)=34.61, p=0.01$].

Multiple paired comparisons, with Bonferroni correction, were run between the converted scores obtained in the various conditions. The Bonferroni corrected statistical significance level was set at $p < 0.017$ ($\alpha=0.05$). The results are shown in Table II. The comparisons indicated no statistically significant differences between the HAR and EAS scores at

TABLE II. Multiple paired comparisons between the converted scores obtained in the various conditions for EAS processing. ** indicates significantly higher, with Bonferroni corrected $p < 0.017$, $\alpha=0.05$.

	-1 dB	1 dB	3 dB	5 dB
HAR vs. EAS				
HAR+PAR vs. EAS	**	**	**	**
HAR+PAR vs. HAR	**	**	**	**

all three SNR levels, suggesting that regenerating the main harmonics alone cannot provide benefits with EAS processing for telephone speech in steady-state noise. The scores obtained with the HAR+PAR stimuli were significantly higher than those obtained with EAS at various SNR levels, suggesting that regenerating the signal's main harmonics and the neighboring partials below 600 Hz can improve the benefits with EAS processing for telephone speech in steady-state noise. Figure 6 also shows pilot data collected with three subjects for the perception of EAS-vocoded clean telephone speech, and there appears to be a ceiling effect for the three processing conditions.

III. EXPERIMENT 2: EFFECTS OF INTRODUCING LOW-FREQUENCY INFORMATION PRIOR TO VOCODER PROCESSING

In the previous experiment, we assessed the potential benefit of introducing low-frequency (0–600 Hz) harmonics in EAS-vocoded telephone speech. The target population for the proposed speech coding algorithm are cochlear implant users with preserved residual hearing in the implanted ear. A short electrode array (e.g., 10 mm in length) is often inserted in those patients with the aim of preserving the residual hearing in the low frequencies (Gantz *et al.*, 2005). The present experiment considers a different target population, namely CI users with no residual hearing in either ear (for most of these CI users, the electrode array does not reach the apical area of the cochlea. However, what is typically done is that the analysis filters capture low-frequency information as low as 200–300 Hz, and that information is output to the most apical electrode which in some cases might be at the 800-Hz place. Consequently, the acoustic information is spectrally up-shifted). Hence, unlike those listeners tested in Exp. 1, the listeners in the present study do not have access to low-frequency acoustic information, only low-frequency vocoded information. According to our prior study (Hu and Loizou, 2010), combined with noise reduction techniques, the harmonics regeneration concept can provide significant intelligibility benefits when used prior to vocoder processing. Unlike the prior study, however, the present experiment uses vocoded speech processed via the telephone network (300–3400 Hz), thus containing no useful information in the most apical channels.

A. Methods

A different set of eleven normal-hearing native speakers of American English participated in this experiment. All subjects were paid for their participation, and all of them were undergraduate and graduate students at the University of

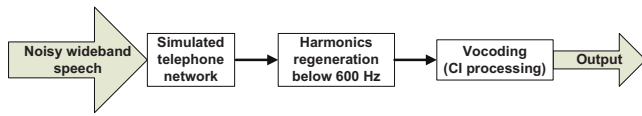


FIG. 7. (Color online) Block diagram of the overall system for simulated CI processing.

Texas-Dallas. The target speech materials and the masker were the same as in Experiment 1, and the filtered speech-shaped noise was added to the filtered speech signal at 5, 10 and 15 dB. The degraded speech signals were downsampled to 8 kHz.

The experiments were designed to evaluate the benefits of harmonics regeneration when used in a pre-processing stage to vocoder (cochlear implant) processing. Figure 7 shows the block diagram of the overall system. A total of three processing conditions were used for this purpose. The first condition was designed to simulate cochlear implant processing. As the first step, a pre-emphasis filter with 2000 Hz cutoff frequency and 3 dB/octave rolloff was applied to the signal. An eight-channel noise-excited vocoder was utilized (Shannon *et al.*, 1995). The telephone speech signal was bandpassed into eight frequency bands between 80 Hz and 3400 Hz using sixth-order Butterworth filters. For the specified frequency range, the ERB filter spacing (Glasberg and Moore, 1990) was used to allocate the eight frequency channels (see Table III). The envelopes of the bandpassed signals were obtained by full-wave rectification followed by low-pass filtering using a second-order Butterworth filter with a 400 Hz cutoff frequency. The extracted temporal envelopes were modulated with white noise, and bandpass filtered through the same analysis bandpass filters. The resulting (narrow-band filtered) waveforms in each channel were finally summed to generate the vocoded stimuli. The level of the synthesized speech signal was scaled to have the same root mean square (RMS) value as the original speech signal. We denote this condition as V.

The other two processing conditions were the same as in Experiment 1. More precisely, the HAR and HAR+PAR stimuli were first constructed using the method described in Sec. II B. These stimuli were subsequently vocoded (as described above) to generate a new set of stimuli, which we denote as vHAR and vHAR+PAR respectively. Note that the vocoding of the HAR and HAR+PAR stimuli resulted in three new envelopes in the first three channels spanning the

TABLE III. Low and high cutoff frequencies (at -3 dB) for the 8 channels used in the vocoding stage.

Channel	Low (Hz)	High (Hz)
1	80	191
2	191	343
3	343	549
4	549	830
5	830	1212
6	1212	1731
7	1731	2438
8	2438	3400

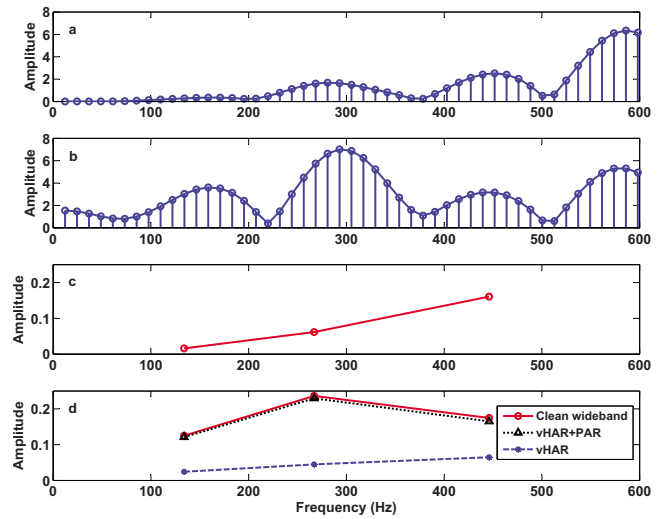


FIG. 8. (Color online) (a) FFT magnitude spectrum of a voiced segment extracted from telephone-processed speech. (b) FFT magnitude spectrum of a voiced segment extracted from wideband speech. (c) Envelopes of telephone speech, shown in panel (a), for the first three most-apical channels. (d) The vHAR and vHAR+PAR envelopes compared against the envelopes obtained from wide band speech.

frequency range of 80–549 Hz (see Table III). The remaining five channels (>549 Hz) contained vocoded signals constructed using the method described above. Note that the vocoded signals (in channels with center frequency >549 Hz) were the same in Exp. 1 and Exp. 2. Figure 8 shows example envelopes of the vHAR and vHAR+PAR stimuli. Panel b shows the FFT magnitude spectrum of a voiced segment extracted from wideband clean speech. The “clean wideband” in panel d is obtained from the segment shown in panel b. vHAR is obtained by integrating the energy in the 3 low-frequency bands ([80–191], [191–343], [343–549]) of the signal shown in panel c of Fig. 4. The frequency location of the main harmonics relative to the (low and high) edges of each band can affect the amount of energy accumulated in each band. Similarly, vHAR+PAR was obtained by integrating the energy within the same 3 bands for the signal shown in panel d of Fig. 4. From panel d of Fig. 8, it becomes clear that the envelopes of the vHAR+PAR stimuli approximate well the envelopes extracted from the clean wideband signal.

B. Procedure

The procedure was identical to Experiment 1 except that, to familiarize each subject with the vocoded stimuli, a training session was administered prior to the formal testing, and each subject listened to vocoded speech stimuli for about 15–20 min. In total, there were 9 testing conditions ($=3$ SNR levels $\times 3$ processing methods).

C. Results

The mean percent correct scores for all conditions are shown in Fig. 9. Performance was measured in terms of percent of words identified correctly (all words were scored). The scores were first converted to rational arcsine units (RAU) using the rationalized arcsine transform proposed by

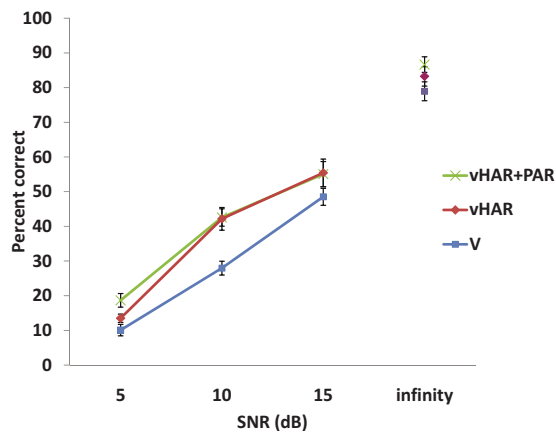


FIG. 9. (Color online) Mean percent correct scores as a function of SNR level. The error bars denote ± 1 standard error of the mean.

Studebaker (1985). To examine the effect of SNR level and type of harmonic structure (main harmonics only vs. main harmonics plus neighboring partials, i.e., vHAR vs. vHAR+PAR) regenerated, we subjected the RAU-converted scores to statistical analysis using the converted score as the dependent variable, and the SNR level and type of harmonic structure as the two within-subjects factors. Analysis of variance (ANOVA) with repeated measures indicated significant effects of SNR level [$F(2,20)=111.71, p < 0.0005$] and type of harmonic structure [$F(2,20)=61.73, p < 0.0005$]. There were no significant between-factor interactions between SNR level and type of harmonic structure [$F(4,40)=56.07, p = 0.177$].

Multiple paired comparisons, with Bonferroni correction, were run between the converted scores obtained in the various conditions. The Bonferroni corrected statistical significance level was set at $p < 0.017$ ($\alpha = 0.05$). The results are shown in Table IV. The comparisons indicated no statistically significant differences between the vHAR and vHAR+PAR scores at all three SNR levels. The scores obtained in the vHAR+PAR condition were significantly higher than those obtained in the control V condition at low SNR levels (5 and 10 dB), suggesting that regenerating the signal's main harmonics and the neighboring partials below 600 Hz can improve the intelligibility of vocoded telephone speech in steady-state noise at low SNR levels. Figure 9 also shows pilot data for the perception of CI-vocoded clean telephone speech.¹

IV. GENERAL DISCUSSION AND CONCLUSIONS

Using the harmonics-regeneration techniques introduced in Hu (2010) and Hu and Loizou (2010), the present study

TABLE IV. Multiple paired comparisons between the converted scores obtained in the various conditions for Vocoder processing. ** indicates significantly higher, with Bonferroni corrected $p < 0.017$, $\alpha = 0.05$.

	5 dB	10 dB	15 dB
vHAR vs. V		**	
vHAR+PAR vs. V	**	**	
vHAR+PAR vs. vHAR			

examined the potential benefits of regenerating low-frequency (< 600 Hz) harmonics in the perception of EAS- and CI-vocoded telephone speech. The two simulation experiments attempted to answer the question whether telephone speech recognition in noise can be improved partly by introducing, prior to EAS and CI processing, low-frequency information encompassing the missing (due to the telephone network) information residing in the 0–300 Hz band. The present study focused on extending the bandwidth toward the low frequency range, which is in contrast to previous studies that extended the bandwidth toward the higher frequency range (> 3.4 kHz).

Although the present study used similar harmonics-regeneration techniques to those in Hu (2010) and Hu and Loizou (2010), there exist some differences and similarities between them in terms of outcomes. The slight differences in outcomes were due to the fact that the harmonics-regeneration technique was applied in the prior studies in the context of noise reduction, which is not considered here. The same upper cutoff frequency (600 Hz) was used in Hu (2010) to regenerate the main harmonics and adjacent partials. Results demonstrated that compared to noise reduction alone, further regenerating the main harmonics for EAS processing significantly improved speech recognition in noise in the low SNR conditions but not in the high SNR conditions. After regenerating the main harmonics along with the adjacent partials, speech recognition in noise was significantly improved in both low and high SNR conditions. In contrast, in the present study, introducing only the main harmonics below 600 Hz did not improve telephone speech recognition at all SNR levels in EAS processing, but introducing the main harmonics along with the adjacent partials below 600 Hz significantly improved telephone speech recognition at all SNR levels (Exp. 1). In the context of noise reduction in CI processing (Hu and Loizou, 2010), when a higher upper cutoff frequency (1 or 3 kHz) was used, preserving the main low-frequency harmonics (spanning 1 or 3 kHz) alone was not found to be beneficial. Preserving, however, the main harmonics along with the adjacent partials was found to be critically important and yielded substantial improvements in intelligibility. In the present study, based on an upper cutoff frequency of 600 Hz, introducing the main harmonics along with the adjacent partials was only found to be beneficial at the low SNR levels in CI processing (Exp. 2), and the benefits of introducing only the main harmonics were limited. In brief, our prior studies and the present study yielded one consistent outcome: a larger benefit in intelligibility is introduced when the main harmonics are preserved along with adjacent partials, compared to the situation where only the main harmonics are preserved.

A. Perception of EAS-vocoded telephone speech

The data in Exp. 1 showed that performance in the HAR condition did not differ significantly from the performance in the control EAS condition. It should be noted that unlike other studies [e.g., Brown and Bacon (2009b)], the control EAS condition did not contain low-frequency (< 300 Hz) acoustic information due to the limited telephone bandwidth.

In the HAR condition, subjects had access to the clean harmonics, spanning the range of 0–600 Hz, along with vocoded information in the higher frequencies. Yet, despite having access to reliable F0 cues, subjects did not show any benefit in intelligibility. In some respects², this outcome bears similarity with that reported by [Brown and Bacon \(2009a\)](#). In that study, a single amplitude and frequency modulated tone was used in place of the low-frequency acoustic information, and was found to yield numerically lower performance than the EAS condition³ (73% for EAS and 64% with the acoustic tone supplemented with electric information, although not significantly different), at least at an average SNR level of 12 dB. The data from Exp. 1 suggest that no differences are to be expected if only main harmonics are used. In our case, as many as 6 harmonics were introduced in the 0–600 Hz range, assuming a typical F0 value of a male talker (F0=100 Hz). This suggests that the F0 cues present in the low frequencies (<600 Hz) might not be reliable or perhaps salient when only the main harmonics are present, at least in the steady-background noise tested in this study. On the other hand, a significant boost (20%) in performance was obtained in the HAR+PAR condition wherein low-frequency harmonics along with adjacent partials (see Fig. 6) were introduced. We speculate that the spectral information contained in the HAR+PAR stimuli was more salient and easier for listeners to access compared with the rather sparse spectral information contained in the HAR stimuli (see Fig. 4). We suspect that the presence of partials in the HAR+PAR condition made the stimuli more natural and perhaps easier to fuse with the vocoded portion. This enabled listeners to extract and utilize F0/F1 and associated voicing information more effectively. Such information, when integrated with the vocoded information in the higher frequencies, is critically important as it can facilitate better word/syllable segmentation and glimpsing of the target (during voiced segments).

A glimpsing account for the benefit of EAS in noise was proposed by [Kong and Carlyon \(2007\)](#) and [Li and Loizou \(2008b\)](#). In the latter study, two factors were posited to play a critical role in receiving EAS benefit when LP information is supplemented with higher-frequency vocoded information: ability to detect glimpses and ability to integrate the glimpsed information. In the [Li and Loizou \(2008b\)](#) study, the listeners were able to glimpse the target information during the temporal dips of the competing talker. In contrast, in the present study, the listeners were able to glimpse the target during the voiced speech segments of the utterance, as the HAR+PAR stimuli approximated well the clean LP stimuli (see Fig. 4).

Since the HAR+PAR stimuli were constructed only during the voiced speech segments (unvoiced segments remained noise masked), accurate vowel/consonant boundaries were available to the listeners. That is, voicing information was reliable in the HAR+PAR condition (voicing information was also present in the HAR condition, but listeners were not able to utilize that information as effectively). As reported in previous studies ([Li and Loizou, 2008a](#); [Spitzer et al., 2009](#); [Zhang et al., 2010](#)), having access to accurate voicing information can facilitate word/syllable segmenta-

tion and enable listeners to identify more words in the noisy speech stream. In the study by [Li and Loizou \(2008a\)](#), a 15% improvement in intelligibility was obtained at low SNR levels when the listeners had access to the low-frequency region (0–500 Hz) of the clean obstruent consonant spectra in otherwise masked sentences (sonorant segments were left corrupted). This improvement was attributed to better transmission of voicing information and enhanced access to acoustic landmarks, such as those evident in spectral discontinuities signaling the onsets/offsets of weak consonants (e.g., /t/) and vowels. These landmarks are posited to be critically important for understanding speech in noise as they aid listeners to better determine the syllable structure and word boundaries ([Stevens, 2002](#)). In contrast to the listeners in [Li and Loizou \(2008a\)](#), the listeners in the present study had better access to the voiced (rather than unvoiced) segments of the utterance, but in both studies the vowel/consonant boundaries were clear. [Zhang et al. \(2010\)](#) also pointed out the importance of voicing as a landmark to aid lexical access for EAS patients. Aside from voicing information, which can be conveyed by F1 as well as other cues (e.g., vowel duration preceding final-syllable stops), the natural F0 contours present in the acoustic portion can also facilitate lexical segmentation in EAS users. [Spitzer et al. \(2009\)](#) showed that to guide lexical segmentation, syllabic stress cues from F0 contours were used by both CI and EAS listeners.

In brief, the data from Exp. 1 as well as the outcomes from previous studies ([Qin and Oxenham, 2006](#); [Kong and Carlyon, 2007](#); [Li and Loizou, 2008b](#); [Brown and Bacon, 2009b](#); [Spitzer et al., 2009](#); [Zhang et al., 2010](#)) suggest that the benefit introduced with the HAR+PAR stimuli in noisy conditions can be attributed to the fact that listeners had reliable access to a combination of F0, glimpsing and lexical segmentation cues.

B. Perception of CI-vocoded telephone speech

The data in Exp. 2 showed that performance in the vHAR condition did not differ significantly from performance in the vHAR+PAR condition. The stimuli in both conditions yielded significant improvements in intelligibility over that obtained using the unprocessed stimuli, at least at the low SNR levels. Unlike the low-frequency acoustic information present in the stimuli in Exp. 1, the information present in the stimuli in Exp. 2 was vocoded. For that reason, we believe that the listeners in Exp. 2 were not able to glimpse as effectively the target during the voiced segments of the utterance. We attribute the improvement in performance obtained with the vHAR+PAR (and vHAR) stimuli to the improved transmission of voicing information and better access to lexical segmentation cues. The vHAR stimuli provided accurate voicing information, despite the fact that these stimuli did not approximate well the clean envelopes (see Fig. 8). We therefore attribute the improvement in performance with the vHAR stimuli to reliable voicing information. As mentioned earlier, a 15% improvement in intelligibility was obtained in the study by [Li and Loizou \(2008a\)](#) when the listeners had access to low-frequency (0–500 Hz)

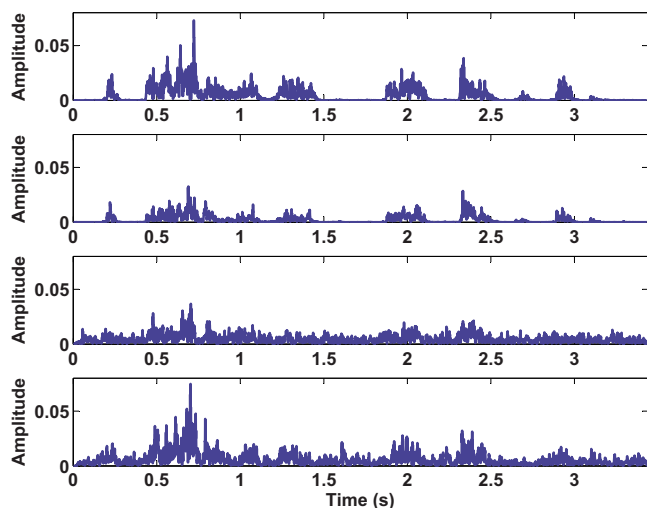


FIG. 10. (Color online) An example plot of the vocoded speech at channel 2 (center frequency=267 Hz). The top panel shows the temporal envelope of the vocoded wideband clean speech. The second panel shows the temporal envelope of the vocoded clean telephone speech. The third panel shows the temporal envelope of the vocoded noisy telephone speech at SNR = 0 dB. The bottom panel shows the temporal envelope of the vocoded speech in the vHAR+PAR condition which preserved the main harmonics along with the adjacent partials.

voicing information. This level of improvement is consistent with the improvement (12%) observed in Exp. 2.

Figure 10 shows an example plot of the vocoded speech at channel 2 (center frequency=267 Hz) for the clean wideband signal, clean telephone signal, noisy telephone signal at 0 dB, and vHAR+PAR processed signal. As can be seen in the bottom panel, the temporal envelope of the vHAR+PAR provides clear evidence of the vowel/consonant boundaries, which are critically important for lexical access (Stevens, 2002). In contrast, the vowel/consonant boundaries in the telephone vocoded signal (panel c) at 0 dB SNR are faint and for the most part are missing. In the context of cochlear implants, the vowel/consonant boundaries in vocoded speech are blurred in noisy environments, and this can be attributed to two reasons: envelope compression and reduced dynamic range (Li and Loizou, 2009). In current CI systems, the envelopes extracted from each band are compressed with a logarithmic function in order to map the wide acoustic dynamic range to the small (5–15 dB) electrical dynamic range. This envelope compression smears the acoustic landmarks a great deal (more so in noise) making it extremely difficult for CI users to identify word boundaries. The proposed harmonics regeneration technique enhances access to the low-frequency acoustic landmarks evident in vowel/consonant boundaries.

No significant benefit in intelligibility was observed at high SNR levels (15 dB) or in quiet conditions, when low-frequency information was introduced prior to vocoder processing. This suggests that listeners must be using different cues when the masker level is low (or absent). Or, this could mean that at high SNR levels the vowel/consonant boundaries are already clear in the temporal envelopes, hence making the boundaries more evident to the listeners provides no additional benefit. In brief, introducing low-frequency harmonic information prior to vocoder processing can bring sig-

nificant benefits to the intelligibility of vocoded telephone speech at low SNR levels. This benefit is attributed primarily to the listeners having better access to voicing and lexical segmentation cues.

C. Practical implications

The present study used F0 values and harmonic amplitudes extracted from clean wideband speech; hence the results reported here reflect the full potential of the proposed algorithms. In realistic scenarios, the F0 values need to be estimated from telephone speech using F0 detection techniques. Several such techniques are currently available, with some that have been proved to be robust (e.g., Wang and Seneff (2000)). To estimate the harmonic amplitudes from noisy telephone speech, frequency-equalization techniques, such as those used in Terry *et al.* (1992), can be applied first to compensate for the low-frequency characteristics of the telephone filter. The harmonic amplitudes can then be estimated using codebook-based techniques that capitalize on the correlation between harmonics (Chu, 2004; Zavarehei *et al.*, 2007) or can be obtained using nonlinear functions (Plapous *et al.*, 2006) or adaptive comb filtering techniques (Nehorai and Porat, 1986). Following the estimation of the harmonic amplitudes, the adjacent partials can be easily generated using the proposed Gaussian-shaped function.

In brief, extending the missing low-frequency region (<300 Hz) of telephone speech seems to be an easier task than extending the high-frequency (>3.4 kHz) region of the spectrum. For a typical F0 value of a male talker (F0 = 100 Hz), for instance, only 7 parameters (=F0+6 harmonic amplitudes) need to be estimated from telephone speech. These 7 parameters can then be used to regenerate low-frequency information spanning the range of 0–600 Hz. As demonstrated in the present study, significant benefits in telephone speech intelligibility can be obtained when listeners have access to low-frequency information in noisy situations.

ACKNOWLEDGMENTS

This research was supported by NIH/NIDCD Grant Nos. R03-DC008887 (Y.H.) and R01-DC07527 (P.C.L.). The authors thank Dr. Magdalena Wojtczak, Christopher A. Brown and the anonymous reviewer for their very helpful comments, and Jinesh Jain for his assistance with data collection.

¹Six subjects were tested and paired comparisons with Bonferroni correction showed no statistical significance between the three processing conditions.

²There are a few differences between the conditions in the study by Brown and Bacon (2009a) and the conditions in our study. First, in their study, the EAS condition did not contain noise in the acoustic portion (noise was only present in the electric portion). In contrast, in our study noise was present in both acoustic (band-limited) and vocoded portions. Second, the stimuli were presented in Brown and Bacon (2009a) at different SNR levels for each subject. In contrast, in our study the same SNR levels (–1, 1, 3 or 5 dB) were used for all subjects. Despite these differences, there is similarity between the HAR+PAR stimuli of this study and their EAS-tone stimuli, which were synthesized by replacing the acoustic portion with an amplitude-modulated tone. The HAR+PAR stimuli used in our study consisted of multiple (rather than a single) harmonics with adjacent partials, and the EAS-tone stimuli used in Brown and Bacon (2009a) can

be approximately viewed as a special case of our HAR+PAR stimuli, when only the first harmonic along with the adjacent partials are retained (the partials in Brown and Bacon (2009a) have an approximately $32 (=16 \times 2)$ Hz bandwidth compared with 100 Hz in the present study).

³It should be noted that in the study by Brown and Bacon (2009a) the control EAS condition contained low-frequency information spanning 0–600 Hz, whereas in the present study the stimuli used in the EAS condition did not contain low-frequency (<300 Hz) acoustic information due to the limited telephone bandwidth.

- Brown, C. A., and Bacon, S. P. (2009a). "Achieving electric-acoustic benefit with a modulated tone," *Ear Hear.* **30**, 489–493.
- Brown, C. A. and Bacon, S. P. (2009b). "Low-frequency speech cues and simulated electric-acoustic hearing," *J. Acoust. Soc. Am.* **125**, 1658–1665.
- Chu, W. C. (2004). "Vector quantization of harmonic magnitudes in speech coding applications—A survey and new technique," *EURASIP J. Appl. Signal Process.* **2004**, 2601–2613.
- Cohen, N. L., Waltzman, S. B., and Shapiro, W. H. (1989). "Telephone speech comprehension with use of the nucleus cochlear implant," *Ann. Otol. Rhinol. Laryngol. Suppl.* **142**, 8–11.
- Cray, J. W., Allen, R. L., Stuart, A., Hudson, S., Layman, E., and Givens, G. D. (2004). "An investigation of telephone use among cochlear implant recipients," *Am. J. Audiol.* **13**, 200–212.
- Dorman, M. F., Gifford, R. H., Spahr, A. J., and McKarns, S. A. (2008). "The benefits of combining acoustic and electric stimulation for the recognition of speech, voice and melodies," *Audiol. Neuro-Otol.* **13**, 105–112.
- Fu, Q.-J., and Galvin, J. J. (2006). "Recognition of simulated telephone speech by cochlear implant users," *Am. J. Audiol.* **15**, 127–132.
- Gantz, B. J., Turner, C. W., and Gfeller, K. E. (2006). "Acoustic plus electric speech processing: Preliminary results of a multicenter clinical trial of the Iowa/Nucleus hybrid implant," *Audiol. Neuro-Otol.* **11**, 63–68.
- Gantz, B. J., Turner, C. W., Gfeller, K. E., and Lowder, M. W. (2005). "Preservation of hearing in cochlear implant surgery: Advantages of combined electrical and acoustical speech processing," *Laryngoscope* **115**, 796–802.
- Glasberg, B., and Moore, B. (1990). "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.* **47**, 103–138.
- Hong, M.-J., Chen, H.-C., Hsu, C.-J., and Fu, Q.-J. (2007). "Telephone speech perception by Mandarin-speaking cochlear implantees," *Ear Hear.* **28**, 665–695.
- Hu, Y. (2010). "A simulation study of harmonics regeneration in noise reduction for electric and acoustic stimulation," *J. Acoust. Soc. Am.* **127**, 3145–3153.
- Hu, Y., and Loizou, P. (2010). "On the importance of preserving the harmonics and neighboring partials prior to vocoder processing: Implications for cochlear implants," *J. Acoust. Soc. Am.* **127**, 427–434.
- IEEE (1969). "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.* **17**, 225–246.
- Ito, J., Nakatake, M., and Fujita, S. (1999). "Hearing ability by telephone of patients with cochlear implants," *Otolaryngol.-Head Neck Surg.* **121**, 802–804.
- Kepler, L., Terry, M., and Sweetman, R. H. (1992). "Telephone usage in the hearing-impaired population," *Ear Hear.* **13**, 311–330.
- Kondoz, A. M. (1999). *Digital Speech: Coding for Low Bit Rate Communication Systems* (Wiley, New York), pp. 57–84.
- Kong, Y., and Carlyon, R. P. (2007). "Improved speech recognition in noise in simulated binaurally combined acoustic and electric stimulation," *J. Acoust. Soc. Am.* **121**, 3717–3727.
- Kong, Y., Stickney, G. S., and Zeng, F. (2005). "Speech and melody recognition in binaurally combined acoustic and electric stimulation," *J. Acoust. Soc. Am.* **117**, 1351–1361.
- Li, N., and Loizou, P. C. (2008a). "The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise," *J. Acoust. Soc. Am.* **124**, 3947–3958.
- Li, N., and Loizou, P. C. (2008b). "A glimpsing account for the benefits of simulated combined acoustic and electric hearing," *J. Acoust. Soc. Am.* **123**, 2287–2294.
- Li, N., and Loizou, P. C. (2009). "Factors affecting masking release in cochlear implant vocoded speech," *J. Acoust. Soc. Am.* **126**, 338–348.
- Liu, C., Fu, Q.-J., and Narayanan, S. S. (2009). "Effect of bandwidth extension to telephone speech recognition in cochlear implant users," *J. Acoust. Soc. Am.* **125**, EL77–EL83.
- Loizou, P. (2007). *Speech Enhancement: Theory and Practice* (CRC, Boca Raton, FL), pp. 589–592.
- Milchard, A. J., and Cullington, H. E. (2004). "An investigation into the effect of limiting the frequency bandwidth of speech on speech recognition in adult cochlear implant users," *Int. J. Audiol.* **43**, 356–362.
- Moore, B., and Glasberg, B. (1983). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.* **74**, 750–753.
- Nehorai, A., and Porat, B. (1986). "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Process.* **34**, 1124–1138.
- P.862 (2000). Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs (ITU-T Recommendation P.862).
- Parikh, G., and Loizou, P. C. (2005). "The influence of noise on vowel and consonant cues," *J. Acoust. Soc. Am.* **118**, 3874–3888.
- Plapous, C., Marro, C., and Scalart, P. (2006). "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.* **14**, 2098–2108.
- Qin, M. K., and Oxenham, A. J. (2006). "Effects of introducing unprocessed low-frequency information on the reception of envelope-vocoder processed speech," *J. Acoust. Soc. Am.* **119**, 2417–2426.
- Shannon, R., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- Spitzer, S., Liss, J., Spahr, A. J., Dorman, M. F., and Lansford, K. (2009). "The use of fundamental frequency for lexical segmentation in listeners with cochlear implants," *J. Acoust. Soc. Am.* **125**, EL236–EL241.
- Stevens, K. N. (2002). "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Am.* **111**, 1872–1891.
- Studebaker, G. A. (1985). "A "rationalized" arcsine transform," *J. Speech Hear. Res.* **28**, 455–462.
- Terry, M., Bright, K., Durian, M., Kepler, L., Sweetman, R. H., and Grim, M. (1992). "Processing the telephone speech for the hearing impaired," *Ear Hear.* **13**, 70–79.
- Turner, C. W., Gantz, B. J., Vidal, C., Behrens, A., and Henry, B. A. (2004). "Speech recognition in noise for cochlear implant listeners: Benefits of residual acoustic hearing," *J. Acoust. Soc. Am.* **115**, 1729–1735.
- Wang, C., and Seneff, S. (2000). "Robust pitch tracking for prosodic modeling in telephone speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1343–1346.
- Zavarehei, E., Vaseghi, S., and Yan, Q. (2007). "Noisy speech enhancement using harmonic-noise model and codebook-based post-processing," *IEEE Trans. Audio, Speech, Lang. Process.* **15**, 1194–1203.
- Zhang, T., Dorman, M. F., and Spahr, A. J. (2010). "Information from the voice fundamental frequency (F0) region accounts for the majority of the benefit when acoustic stimulation is added to electric stimulation," *Ear Hear.* **31**, 63–69.