Gender and speaker identification as a function of the number of channels in spectrally reduced speech

Julio Gonzalez[a)]
Departmento de Psicologia Basica, Clinica y Psicobiologia, Universidad Jaume I, Castellon, 12071 - Castellon, Spain

Juan C. Oliver[b)]
Departmento de Psicologia Evolutiva, Educativa, Social y Metodologia, Universidad Jaume I, Castellon, 12071 - Castellon, Spain

(Received 8 August 2003; revised 26 January 2005; accepted 30 March 2005)

Considerable research on speech intelligibility for cochlear-implant users has been conducted using acoustic simulations with normal-hearing subjects. However, some relevant topics about perception through cochlear implants remain scantly explored. The present study examined the perception by normal-hearing subjects of gender and identity of a talker as a function of the number of channels in spectrally reduced speech. Two simulation strategies were compared. They were implemented by two different processors that presented signals as either the sum of sine waves at the center of the channels or as the sum of noise bands. In Experiment 1, 15 subjects determined the gender of 40 talkers (20 males + 20 females) from a natural utterance processed through 3, 4, 5, 6, 8, 10, 12, and 16 channels with both processors. In Experiment 2, 56 subjects matched a natural sentence uttered by 10 talkers with the corresponding simulation replicas processed through 3, 4, 8, and 16 channels for each processor. In Experiment 3, 72 subjects performed the same task but different sentences were used for natural and processed stimuli. A control Experiment 4 was conducted to equate the processing steps between the two simulation strategies. Results showed that gender and talker identification was better for the sine-wave processor, and that performance through the noise-band processor was more sensitive to the number of channels. Implications and possible explanations for the superiority of sine-wave simulations are discussed. [DOI: 10.1121/1.1928892]

PACS number(s): 43.71.Bp, 43.71.Ky, 43.72.Ar, 43.66.Ts [KWG] Pages: 461–470

I. INTRODUCTION

Speech is a robust signal that is resistant to many forms of information reduction. Speech recognition does not require all the fine spectral information present in the natural signal. This circumstance allows that deaf individuals fitted with cochlear implants can understand speech through a relatively small number of electrodes, or channels.

A useful approach in research is to test normal-hearing listeners with speech signals that have been processed in the manner of a cochlear-implant (CI) signal processor. In these types of experiments, two main signal processors have been used to create simulations of cochlear-implant signals. In the strategy of Shannon *et al.* (1995), the speech signal is bandpass filtered into *n* bands, or channels, and the envelope of each band is extracted, and used to modulate white noise, which is band limited with the same bandpass filter. This way, signals are presented as the sum of noise bands whose bandwidths were equal to the bandwidth of the original analysis channels. These authors showed that high levels of understanding of speech produced in quiet could be achieved using as few as four bands. Dorman *et al.* (1997) synthesized simulations as a sum of *n* sine waves at the center of the channels rather than noise bands. As in Shannon *et al.* (1995), high level of speech understanding could be achieved using only four channels. This strategy was based on the observation that CI users usually report after individual channel stimulation that the signals sound like beep tones and not like bands of noise. Comparison between both processors showed that the nature of the output signal, either noise bands or sine waves, made only a small difference in speech intelligibility (Dorman *et al.* 1997). Results also showed that the number of channels needed to approach asymptotic performance varied with the difficulty of the speech material. For the most difficult material (vowels) 8 channels were necessary, whereas for the least difficult material (high-context sentences) 5 channels were sufficient. On the other hand, comparison of speech recognition by CI users and by normal listeners in conditions simulating cochlear implants indicated that both processors work reasonably well as simulations of the CI signal (Dorman *et al.*, 1998; Fu *et al.*, 1998; Loizou *et al.*, 2000; Friesen *et al.*, 2001; Loizou and Poroy, 2001).

To date, most research work on the perception by CI users and simulation studies has been centered on the intelligibility of speech under a wide variety of conditions. Intelligibility is studied using different speech materials (consonants, vowels, words, and sentences) produced by a single or by multiple speakers as a function of the number of channels

[a)]Electronic mail: gonzalez@psb.uji.es
[b)]Electronic mail: oliverr@psi.uji.es

(Dorman *et al.*, 1997; Loizou *et al.*, 1999; Friesen *et al.*, 2001), location of the cutoff frequencies defining the bands (Shannon *et al.*, 1998), misalignments of spectral information (Shannon *et al.*, 1998; Spahr *et al.*, 2002), under different signal-to-noise ratios (Dorman *et al.*, 1998; Friesen *et al.*, 2001; Fu *et al.*, 1998), intensity resolutions (Loizou *et al.*, 1999; Loizou *et al.*, 2000), and other conditions.

However, in the last several years new questions have emerged about the perception through CI devices. For example, there is an increasing interest in sound-direction identification abilities by bilateral-CI users (Hoesel *et al.*, 2002; Hoesel and Tyler, 2003), pitch perception through CI for speech (Hiki and Fukuda, 2000; Qin and Oxeham, 2003) or for music (Gfeller *et al.*, 1997; McDermott and McKay, 1997; Fujita and Ito, 1999; Lobo *et al.*, 2002), timbre recognition (Gfeller *et al.*, 2002), and source identification for familiar environmental sounds (Shafiro *et al.*, 2003).

Perception of the gender and identity of a speaker via acoustic properties of speech is an important issue in natural communication. Many studies show that the acoustic cues for gender and speaker recognition are present in the coarse-grain structure of speech signals, such as the fundamental frequency, formant structure, or the average long-term spectrum (see review of Bricker and Pruzansky, 1976; Wu and Childers, 1991). In some cases, even radically reduced speech signals, such as sine-wave replicas of natural speech formed by two or three pure tones following the formant trajectories, can adequately convey information about the gender and identity of the speaker (Fellowes *et al.*, 1997; Remez *et al.*, 1997; Sheffert *et al.*, 2002). To date, little is known about whether the CI signal has the potential to allow gender and speaker identification. The exploratory work of Cleary and Pisoni (2002) tested discrimination abilities between pairs of female voices in 44 school-age deaf children who had used a CI for at least 4 years. Subjects were asked to answer "same voice" or "different voice." Two conditions were examined: (a) the sentence was held constant across the voices; (b) different sentences were used. In the first condition children performed 68% of correct responses, which although significantly different from chance (50%), suggests that the discrimination task was difficult for them. In the second condition, children were unable to discriminate between unfamiliar speaker's voices (only 57% of success). In some recent preliminary studies, Chinchilla and Fu (2003a, 2003b) provides data on voice gender discrimination from both CI patients and normal-hearing subjects with different simulation strategies and temporal/spectral resolution. Two additional studies provide data on cochlear implant user's ability to discriminate speaker identity (McDonald *et al.*, 2003) and the relative contributions of amplitude and frequency modulations to speaker identification (Kong *et al.*, 2003).

In the present study, the two processors most used in simulation studies were tested for comparative purposes. Normal-hearing adult subjects were tested to assess their ability to recognize the gender and identity of different speakers from acoustic simulations of cochlear-implant signals. We studied signals presented as either the sum of sine waves or as the sum of noise bands with a varying number of frequency channels. In a first experiment, subjects determined the gender of 40 unfamiliar speakers from a natural utterance processed through different numbers of channels with both processors. In the second experiment, listeners matched simulation replicas and natural recordings of a sentence uttered by 10 unfamiliar speakers. In the third experiment, different sentences were used for natural and processed stimuli. Finally, a control fourth experiment was done to equate the processing steps between the two simulation strategies.

II. EXPERIMENT 1: GENDER IDENTIFICATON

A. Method

1. Subjects

15 subjects (10 females and 5 males) with normal speech and audition participated in the experiment. They were students at the University Jaume I of Castellón in Spain with ages ranging from 21 to 30. Subjects participated voluntarily for course credit. None had participated in any other experiment that used CI simulations.

2. Test materials

A Spanish sentence (the question *¿Cuantos años tiene tu primo de Barcelona?* [How old is your cousin from Barcelona?]) was recorded from 40 native speakers of Spanish, 20 males and 20 females, from 25 to 40 years of age. The sentence was uttered at a comfortable level and recorded in a sound-attenuated booth with a Shure SM58 microphone at a distance of about 12 cm from the mouth, and a Sony-TCD D-8 digital audiotape (DAT) recorder with a sample frequency of 44.1 kHz. Then, the voice signal was digitally transferred to a PC computer and converted to 16 bit WAV files. Finally, the files were downsampled to 16 kHz and normalized for overall amplitude.

3. Signal processing

Each natural utterance was processed through a sine-wave and a noise-band processor. The sine-wave processor implementation followed procedures from Loizou *et al.* (1999) and Dorman *et al.* (1997)[11] in the following manner. The signal was first processed through a pre-emphasis filter (1200 Hz high-pass with −6 dB/octave slope) and then bandpassed afterwards into n frequency bands (n=3,4,5,6,8, 10,12,16) using sixth-order Butterworth filters. Following Loizou *et al.* (1999) logarithmic filter spacing was used for $n<8$ and semilogarithmic (mel) filter spacing was used for $n\geqslant 8$ (see the center frequencies and the 3 dB bandwidths of the filters in Tables I and II of Loizou *et al.*, 1999, respectively). The envelope of the signal was extracted by full-wave rectification, and low-pass filtering (second-order Butterworth) with 400 Hz cutoff frequency. Sinusoids were generated with amplitudes equal to the rms energy of the envelopes and with frequencies equal to the center frequencies of the bandpass filters. Finally, the sinusoids of each band were summed and the level was equated to the rms of original.

The noise-band processor was implemented in the following manner based on Shannon *et al.* (1995) and Dorman *et al.* (1997)[2,2]. The signal was first processed through a preemphasis filter (1200 Hz high-pass with −6 dB/octave slope) and was then band-passed into n frequency bands (n=3,4,5,6,8,10,12,16) with the same cut-off frequencies used for the sine-wave processor. Hann-bandpass filtering was performed with a smoothing factor of a tenth of the upper frequency of each band. The envelope of the signal was extracted by full-wave rectification, and by low-pass filtering with 160 Hz cutoff frequency, since Shannon *et al.* (1995) found no difference in performance for low-pass filters set at 160 Hz and above. The envelope of the signal served to modulate white noise, which was band limited with the same Hann-bandpass filter. Finally, each noise band was rescaled to have same power as the original and all the noise bands were summed.

4. Procedure

The experiment was based on a within-subject 2×8 (processors × number of channels) design. Each sentence was assigned to two channel conditions of each processor in a pseudorandom manner, assuring the same number of male and female speakers in each condition. This way, each channel condition for each processor was composed of a fixed set of 10 stimuli (5 male + 5 female speakers). The overall set of stimuli comprised 160 processed sentences derived from the 40 natural samples.

The experiment was performed individually on a Pentium PC equipped with a Creative Labs SoundBlaster 16 soundcard. Each listener completed two series of trials (one per processor) consisting of 80 trials (10 trials × 8 different numbers of channels) in each series. The order of presentation of both series was counterbalanced across the subjects.

Each trial consisted of the presentation of a processed sentence through headphones (AKG model HSC 200) at a comfortable sound level (65–70 dB SPL) and the subject was asked to decide whether the gender of the speaker was male or female. Each trial was auto-administered by the participant. The listeners were not familiar with the speakers whose speech samples were used.

Before each series, subjects were given a practice session with 16 examples of sentences processed through different numbers of channels by the same processor. None of the simulations used in the practice was used in the test. As in Loizou *et al.* (1999), each series followed the same sequential order, starting with the stimuli processed through the largest number of channels (n=16) and ending with stimuli processed by the smaller number of channels (n=3). This sequential design was chosen to give subjects time to adjust for listening to the altered speech signals.

B. Results and discussion

Subjects' decisions were scored as the proportion of correct responses. Results are shown in Fig. 1 (lines sine waves and noise bands). A two-way repeated measures analysis of variance (ANOVA) revealed a main effect of processor type [$F(1,14)=67.50, p<0.001$], with sine-wave processor producing higher scores (mean=0.96) than noise-band processor (mean=0.81). There was also a significant main effect of number of channels [$F(7,98)=23.29, p<0.001$], and a significant interaction between type of processor and number of channels [$F(7,98)=12.46, p<0.001$].

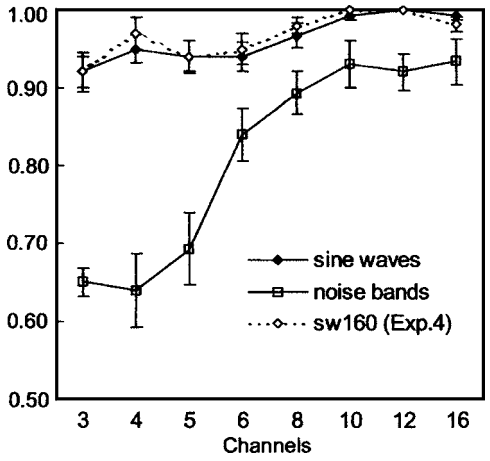

FIG. 1. Experiment 1. Gender identification (proportion correct) as a function of number of stimulation channels. The parameter is processor type: sine-wave output (closed diamonds) or noise-band output (squares). Error bars indicate ± 1 standard error of the mean. Results from Experiment 4 with sine-wave stimuli created with 160 Hz cutoff frequency are also included (sw-160, open diamonds).

The sine-wave processor yielded a high performance in all channel conditions, with a very narrow range of scores from 0.92 (3 channels) to 1 (12 channels). Planned comparisons between adjacent categories (difference contrasts) indicated no statistically significant differences in performance when the number of channels used for this processor was equal or fewer than 8.

Gender identification was worst through the noise-band processor and performance was more sensitive to the number of channels. Scores varied from 0.64 (4 channels) to 0.93 (10 and 16 channels). Planned comparisons between adjacent categories indicated no statistically significant differences in performance when the number of channels used for this processor was equal or fewer than 5. The largest differences were between 5 (0.69) and 6 channels (0.84), and between 6 (0.84) and 8 channels (0.89).

Results (see Fig. 1) showed that gender identification scores for the noise-band processor increased when the number of channels was increased from 4 to 10. However, the sine-wave processor was less sensitive to the number of channels, showing a high performance even at the fewest number of channels. According to the data of Dorman *et al.* (1997), the nature of the output signal, either noise bands or sine waves, makes only a small difference in speech intelligibility. Why did we find a significant difference between both processors in the gender recognition of a speaker? Chinchilla and Fu (2003b) recently studied gender discrimination and vowel recognition by CI and normal-hearing (NH) listeners using sine-wave and noise-band vocoders simulations. Results showed no effect of the simulation type on vowel recognition, but voice gender discrimination was significantly higher through sine-wave than through noise-band simulations.

Recognition of voice gender is dependent upon acoustic information related to the source and vocal tract properties. This information includes fundamental frequency, formant structure, and breathiness (Klatt and Klatt, 1990). Probably, a key factor to account for the superiority of the sine-wave simulations is that sine-wave carriers preserve better than noise carriers some information relevant to the identification of talker gender. This point will be discussed in detail in Sec. VI.

Beyond gender identification, the two following experiments tested the recognition of the identity of a speaker by means of a slightly more complex task.

III. EXPERIMENT 2: SPEAKER IDENTIFICATION—SAME SENTENCE

A. Method

1. Subjects

56 subjects (38 females and 18 males) with normal speech and hearing abilities participated in the experiment. They were students at the University Jaume I of Castellon (Spain), with ages ranging from 20 to 32. Subjects participated voluntarily for course credit. None had taken part in Experiment 1.

2. Test materials

A Spanish sentence (the question *¿Cuantos años tiene tu primo de Barcelona?* [How old is your cousin from Barcelona?]) was recorded from 10 native speakers of Spanish, 5 males and 5 females, with ages ranging from 25 to 39 years. The conditions of recording and creation of WAV files were the same as in Experiment 1.

3. Signal processing

Each natural sentence was treated both by a sine-wave processor and a noise-band processor. Signal processing was the same as in Experiment 1, with the exception that the number of channels used for each processor was $n=3,4,8,16$. The selection of these n followed procedures from the work of Fu *et al.* (1998). All 10 sentences were processed through all the different channel numbers.

4. Procedure

The experiment comprised 8 separate conditions (2 processors $\times$ 4 different numbers of channels). Each subject was randomly assigned to 1 of the 8 conditions (7 subjects per condition).

The experiment was performed in groups of 5 subjects or fewer on Pentium PCs equipped with a Creative Labs SoundBlaster 16 soundcard and the stimuli were individually administered through headphones (AKG model HSC 200) at a comfortable sound level (65–70 dB SPL). The procedure was the same as that used in Remez *et al.* (1997) studying speaker identification from sine-wave replicas. On every trial, a natural sentence was followed by two simulations. One of the pair of simulations was always derived from the natural sentence presented on that trial. The other simulation was derived from one of the other nine sentences (speakers). The subject was asked to report on an answer sheet which of the two simulations was based on the natural sentence presented on each trial.

With 10 different speakers, there were nine comparisons of each simulation with every other one, making 90 trials per condition. The order of the two simulations was counterbalanced along the trials. Because of that, the correct response for half of the trials was "first," and that for the other half was "second." A signal (beep) announced every trial 750 ms before its onset. On each trial, the three stimuli (the natural sentence and the two simulations) were separated by 750 ms of silence. Between each trial, there were 3 s of silence. In every experimental condition, the complete set of 90 trials was administered in blocks of five trials with a short break between blocks. The trials were presented in a pseudorandom order with a maximum of three consecutive similar trials sharing the same correct response.

Before the experimental test, subjects were given a practice session with ten trials of the same condition. None of the stimuli used in the practice were used in the test.

B. Results and discussion

Subjects' decisions were scored as the proportion of correct responses. Results are shown in Fig. 2 (upper panel). A two-way between-subject analysis of variance (ANOVA) revealed a main effect for type of processor [$F(1,48)=85.77, p<0.001$], with the sine-wave processor producing higher scores (mean=0.97) than the noise-band processor (mean=0.83). There was also a significant main effect of number of channels [$F(3,48)=10.56, p<0.001$], but the interaction between type of processor and number of channels did not reach statistical significance [$F(3,48)=1.66, p=0.188$]. However, a separate ANOVA for each processor revealed that the channel variable was not significant for the sine-wave processor [$F(3,24)=1.98, p=0.143$], though it was significant for the noise-band processor [$F(3,24)=10.21, p<0.001$]. Newman–Keuls' *post hoc* tests revealed that performance for the noise-band processor was significantly better with 16 channels than with 4 or 8 channels, which did not differ significantly between them. And performance with 4–8 channels was significantly better than with 3 channels.

The recognition of speaker identity was clearly better with the sine-wave processor. Speaker recognition from stimuli processed through the noise-band processor was more difficult and the number of channels affected performance.

For the purpose of examining the influence of gender on speaker identification, we separated responses to trials formed by stimuli from speakers of the same gender (40 trials per condition), from responses to trials whose stimuli corresponded to speakers of different gender (50 trials per condition). The proportions of correct responses are presented in the lower panel of Fig. 2. A Student t test found a significant effect of the same–different gender variable for the noise-band processor [$t(27)=4.31, p<0.001$], but not for the sine wave processor [$t(27)=0.91, p=0.370$]. As expected, performance in the noise-band processor was better

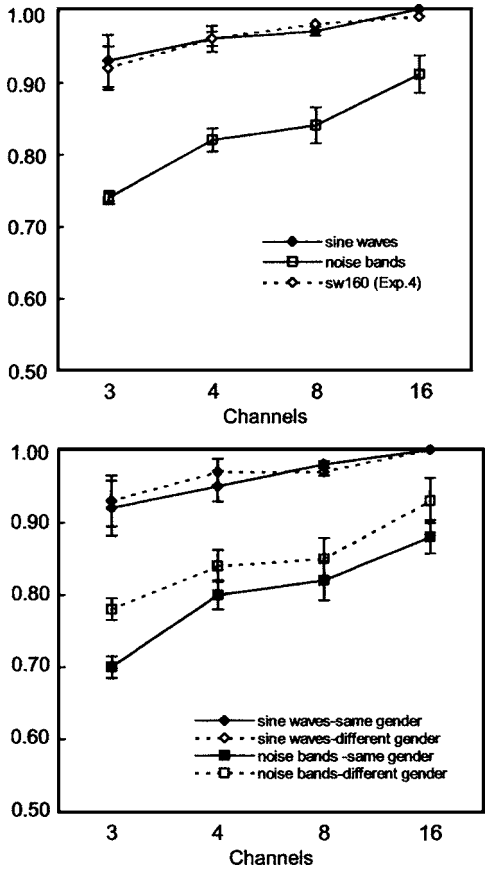

FIG. 2. Upper pannel: speaker identification (proportion correct) as a function of number of stimulation channels from Experiment 2 data (the same sentence was used for natural and processed stimuli). The parameter is processor type: sine-wave output (closed diamonds) or noise-band output (squares). Error bars indicate ± 1 standard error of the mean. Results from Experiment 4 with sine-wave stimuli created with 160 Hz cutoff frequency are also included (sw-160, open diamonds). The lower pannel shows responses to trials whose stimuli were from speakers of the same gender which have been separated from responses to trials whose stimuli were from speakers of different gender.

when the two stimuli to be compared pertained to speakers of different gender. This difference was not found in the sine-wave processor, probably as a result of a ceiling effect. The data from both processor conditions clearly indicated that speaker identification through CI simulations was possible beyond the recognition of gender.

Because sine-wave stimuli have a more regular fine structure, speaker identification could be based on a better modulation detection with the sine-wave carrier than with the noise carrier. In contrast, random level variations in the noise carrier serves to distort the speech envelope. It is probable that less information about speaker identity is available from the noise-band signal. In any case, the superiority of the sinewave simulation will be discussed in more detail in the final section.

On the other hand, it is conceivable that the perceptual judgments in this test may be based on stimulus duration. We reasoned that if the listeners chose between the two simulation alternatives on the basis of a duration strategy, the task would be easier when both simulations were more different in duration, i.e., a positive correlation would emerge across the trials between duration difference and performance. However, no Pearson correlation coefficient was found significant neither of the processor × channel conditions, nor in an average channel condition.

Results of this experiment allow for two alternative explanations. One explanation is that listeners based their performance on the acoustic characteristics of speech of each particular speaker, making an actual speaker identification task. However, given that the natural speech sample of each trial was the model from which the simulation was derived, a different explanation is that listeners based their perceptual judgment on a superficial comparison of the tokens. In this case, subjects would select the correct simulation attending to superficial auditory attributes of specific utterances that are irrelevant to the characteristics of particular speakers. To exclude this possibility we carried out another experiment with the same basic trial procedure using a different natural utterance produced by each talker. This way, listeners who chose the correct simulations would be those who were able to attend to the characteristic acoustic properties of each speaker, beyond the acoustic similarities of specific utterances. This is based on the same logic used by Remez and colleagues (Fellowes *et al.*, 1997; Remez *et al.*, 1997) studying speaker identification from sine wave replicas of speech.

IV. EXPERIMENT 3: SPEAKER IDENTIFICATION—DIFFERENT SENTENCES

A. Method

1. Subjects

A total of 72 subjects (49 females and 23 males) with normal speech and hearing abilities participated in the experiment. They were students at the University Jaume I of Castellon (Spain), with ages ranging from 20 to 32 years. Subjects took part voluntarily for course credit. Thirty-six of them had been participants in Experiment 2, which had been conducted from 3 to 5 months in advance.

2. Test materials

The natural stimuli used in this experiment differed from the utterances that were used as the models for the CI simulations. A new Spanish sentence (the question *¿Vienes mañana al estreno de la película?* [Will you come tomorrow to the opening of the film?]) was recorded from the same 10 speakers (5 males + 5 females) as in Experiment 2. The conditions of recording and creation of WAV files were the same as in Experiments 1 and 2.

The processed stimuli (simulations) were the same as in Experiment 2.

3. Procedure

The experiment comprised eight separate conditions (2 processors × 4 different numbers of channels). Conditions were the same as in Experiment 2: sine-wave processor versus noise-band processor, and number of channels (n=3,4, 8,16). Each subject was randomly assigned to one of the eight conditions (9 subjects per condition).

The experiment was performed in groups of 5 or fewer subjects on Pentium PCs equipped with a Creative Labs SoundBlaster 16 soundcard and the stimuli were individually

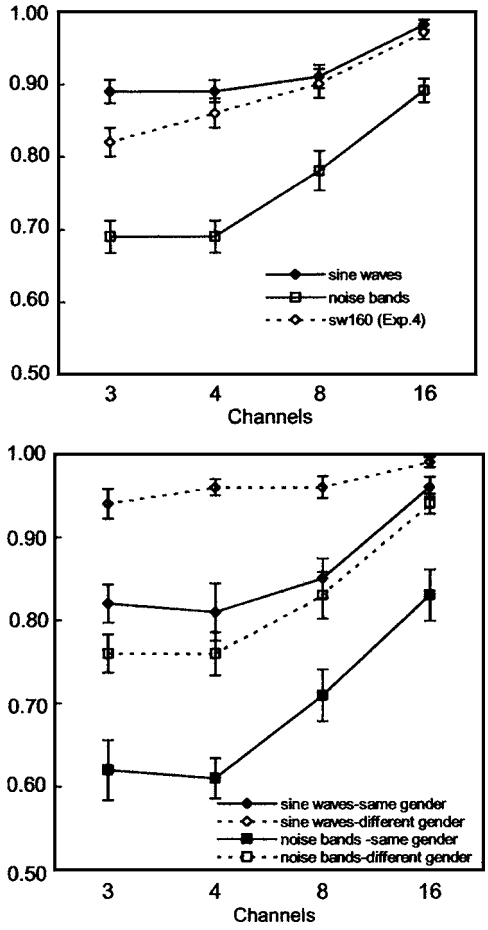

FIG. 3. Upper pannel: speaker identification (proportion correct) as a function of number of stimulation channels from Experiment 3 data (different sentences used for natural and processed stimuli). The parameter is processor type: sine-wave output (closed diamonds) or noise-band output (squares). Error bars indicate ± 1 standard error of the mean. Results from Experiment 4 with sine-wave stimuli created with 160 Hz cutoff frequency are also included (sw-160, open diamonds). The lower pannel shows responses to trials whose stimuli were from speakers of the same gender which have been separated from responses to trials whose stimuli were from speakers of different gender.

administered through headphones (AKG model HSC 200) at a comfortable level (65–70 dB SPL). The procedure was the same as that used in Remez *et al.* (1997). On every trial, a natural sentence (*¿Vienes mañana al estreno de la película?*) was followed by two simulations (both based on *¿Cuantos años tiene tu primo de Barcelona?*). One of the pairs of simulations was always derived from a natural utterance produced by the same speaker who had spoken the natural signal presented on that trial. The other simulation was derived from a natural utterance produced by one of the other nine speakers. The subject was asked to report on an answer sheet which of the two simulations was produced by the same speaker who spoke the natural utterance on each trial.

This experiment followed the same procedure as that in Experiment 2.

B. Results and discussion

The subjects' decisions were scored as the proportion of correct responses. The results are shown in Fig. 3 (upper panel: sine waves and noise bands). A two-way between-subject analysis of variance (ANOVA) revealed a main effect for type of processor $[F(1,64)=135.15, p<0.001]$, with sine-wave processor producing higher scores (mean=0.92) than noise-band processor (mean=0.76). There was also a significant main effect of the number of channels $[F(3,64)=25.26, p<0.001]$, and a significant interaction between type of processor and number of channels $[F(3,64)=3.84, p<0.05]$. Separate ANOVAs for each processor revealed that the channel variable was significant for the sine-wave processor $[F(3,32)=8.40, p<0.001]$, and for the noise-band processor $[F(3,32)=17.17, p<0.001]$. Newman–Keuls' *post hoc* tests identified only two statistically different conditions for the sine-wave processor: 3-4-8, and 16 channels. *Post hoc* tests identified three statistically different conditions for the noise-band processor: 3-4, 8, and 16 channels.

Pooling the data from Experiments 2 and 3, a main experiment effect was found $[F(1,112)=32.9, p<0.001]$, with better performance in Experiment 2 than in 3, as expected. However, no significant interaction was found between type of processor and experiment $[F(1,112)=0.941, p=0.334]$. In Experiment 3 we used stimuli selected to prevent subjects from performing a matching task by listening to acoustic similarities between a natural utterance and its derived processed stimulus. This was accomplished by using as the natural sample of the trials a different sentence from that used as a model to derive the CI simulations. This way, the matching task would be based on more abstract acoustic properties specific to each particular speaker rather than on an exact spectrotemporal match between specific tokens. Logically, performance from Experiment 3 is expected to be more difficult than that from Experiment 2, where in each trial the same token was used both as the natural sample and as the model to derive one of the simulations. Experiment 3 data show two findings:

(1) The sine-wave processor is clearly better than the noise-band processor allowing for speaker identification, showing a good performance (better than 90%) even with the fewest number of channels ($n=3$). The effect of number of channels in the sine-wave processor is only evident when n increases from 8 to 16 channels.
(2) Speaker identification from the noise-band simulations is more difficult but performance is better than chance (50%) in all the channel conditions. Noise-band processor is more sensitive to the number of channels and speaker identification increases from $n=4$ to $n=16$ (see Fig. 3).

As in the previous experiment, we separated in Experiment 3 responses to trials formed by stimuli from speakers of the same gender, from the responses to trials of different gender. The proportions of correct responses are presented in the lower panel of Fig. 3. A Student t test found a significant effect of the same-different gender variable for the sine-wave processor $[t(35)=7.17, p<0.001]$ and for the noise-band processor $[t(35)=8.43, p<0.001]$. As expected, performance in both processors was better when the two stimuli to be compared pertained to speakers of different gender. Never-

theless, data from both processor conditions indicated that speaker identification through CI simulations was possible beyond gender recognition.

V. EXPERIMENT 4: REPLICATION WITH SINE WAVES BASED ON 160 Hz LOW-PASS FILTERING

All the sine-wave stimuli of the present experiments were created using envelope information extracted from each band by low-pass filtering with a 400 Hz cutoff frequency. The rationale behind this cutoff frequency was based on two points. First, we wanted to be coherent with previous studies by using sine-wave processors that conformed to the characteristics of the Med El Corporation's cochlear-implant signal processor (Dorman *et al.*, 1997, 1998; Loizou *et al.*, 1999, 2000). Second, data obtained from intelligibility experiments had shown no difference in performance for low-pass filters set at 160 Hz and above (Shannon *et al.*, 1995). In fact, some experiments comparing speech intelligibility through sine-wave versus noise-band processors used 400 and 160 Hz cutoff frequencies, respectively (Dorman *et al.*, 1997).

Nevertheless, the clear and unexpected superiority of the sine-wave processor in the gender/speaker identification task raises the question whether this performance difference may be at least partially due to the use of a different envelope cutoff frequency for each simulation. To equate for cutoff frequency, we performed a control experiment replicating experiments 1–3 under sine-wave conditions only, using stimuli created with 160 Hz low-pass filtering.

A. Method

1. Subjects

A total of 92 subjects participated in this study. Sixteen subjects replicated Experiment 1 (gender identification). Thirty-six (9×4 channel conditions) replicated Experiment 2 (speaker identification with same sentence). And forty (10×4 channel conditions) replicated Experiment 3 (speaker identification with different sentences). Subjects were students at the University Jaume I of Castellon (Spain) and participated voluntarily for course credit. None of them had participated in any of the previous experiments.

2. Test materials

Only sine waves were used as processed stimuli. Sine-wave stimuli were created in the same way as in Experiments 1–3, except for the cutoff frequency used for low-pass filtering (160 instead of 400 Hz). The number of channels was also the same as in Experiments 1–3. The same natural stimuli from Experiments 2 and 3 were used here in the replication.

3. Procedure

The procedure was the same as in Experiments 1–3.

B. Results and discussion

The subjects' decisions were scored as the proportion of correct responses.

The pattern of results obtained with sine-wave stimuli created with a 160 Hz cutoff frequency (hereafter, sw-160) was the same as that obtained with sine-wave stimuli used in Experiments 1–3, and created with 400 Hz cutoff frequency (hereafter, sw-400), except for the most difficult condition, i.e., speaker identification across different sentences and through 3 channels.

Results are included in all figures adjacent to the findings from the replicated experiments. Gender recognition by means of sw-160 was the same as in the sw-400 condition (see Fig. 1). An analysis of variance (ANOVA) using cutoff frequency as a between-subject factor (sw-160 vs sw-400) showed no main effect $[F(1,29)=0.09, p=0.763]$. On the other hand, comparing performance from sw-160 and noise-band stimuli, gender identification was significantly higher in sw-160 (like sw-400). A between-subject ANOVA (sw-160 versus noise bands) revealed a main effect $[F(1,29)=47.0, p<0.001]$.

Speaker identification under Experiment 2 procedures (stimuli based on the same utterance) is not significantly different for sw-160 in comparison with sw-400 (see Fig. 2, upper panel). An ANOVA (sw-160 vs sw-400) revealed no main effect $[F(1,56)=0.01, p=0.994]$. At the same time, the sw-160 condition yielded higher performance than the noise-band condition: an ANOVA (sw-160 versus noise bands) showed a main effect $[F(1,56)=118.84, p<0.001]$.

It is in the speaker identification under Experiment 3 conditions (stimuli based on different utterances) where a significant difference emerges between results from sw-160 and sw-400 stimuli (see Fig. 3, upper panel). A two-way between-subject ANOVA (with cutoff-frequency and channel as factors) revealed a main effect for cutoff frequency $[F(1,68)=7.90, p<0.01]$, with sw-160 producing lower scores (mean=0.89) than sw-400 stimuli (mean=0.92). There was also a significant main effect of number of channels $[F(3,68)=21.95, p<0.001]$, but the cutoff frequency $\times$ number of channels interaction did not reach statistical significance $[F(3,68)=1.67, p=0.182]$. Comparing cutoff frequencies within each channel condition with a Student t test for independent samples, a significant difference was found between sw-16 and sw-400 only for 3 channels $[t(17)=2.50, p=0.023]$, but not for 4 channels $[t(17)=1.39, p=0.182]$, 8 channels $[t(17)=0.39, p=0.702]$, or 16 channels $[t(17)=1.32, p=0.204]$. A separate ANOVA revealed that the channel variable was significant for the sw-160 processor $[F(3,36)=15.03, p<0.001]$. Newman–Keuls' *post hoc* tests found that performance for sw-160 processor was significantly better with 16 channels than with 3, 4, or 8 channels; and performance with 8 or 4 channels (which did not differ significantly between them) was better than with 3 channels.

In sum, these results show that using the same smoothing filter (160 Hz) for both processors, performance of the sine-wave type remains clearly better that the one of the noise type. Nevertheless, a question remains open. Is it possible that had we used the 400 Hz smoothing filter for noise modulated stimuli the scores would have improved significantly? Noise conditions were more variable than sine-wave conditions. It may be that the 160 Hz filter may interact with

noise to make nearly impossible discerning the periodicity of the speech envelope. Whether performance might have been better under 400 Hz filter and noise carriers is a topic for further research.

VI. GENERAL DISCUSSION

The experiments in the present study assessed the ability of normal-hearing listeners to recognize the gender and identity of a speaker through simulations of cochlear-implant signals. We tested two processors used in simulation studies with a varying number of frequency channels. Previous studies focusing on speech intelligibility had shown that changes in the nature of the output signal from noise bands to sine waves only have a small effect in performance (Dorman *et al.*, 1997).

Nevertheless, when both processors were compared regarding the ability of listeners to recognize the gender and the identity of a speaker, a large and significant difference emerged in all the experiments. Our data showed a substantial superiority of the sine-wave processor in both tasks reaching a very high performance, even in condition with only 3 channels. Research on voice perception has shown that the fundamental frequency of phonation and the spectral properties of natural voice provide strong cues to recognize the gender and identity of a speaker (see review of Bricker and Pruzansky, 1976; Lass *et al.*, 1976, 1980; Klatt and Klatt, 1990; Wu and Childers, 1991; Mullennix *et al.*, 1995; Kreiman, 1997). Listeners take advantage of information provided by the sine-wave simulations that is not present in the noise-band simulations. Acoustic analysis of processed signals from both processors demonstrated that sine-wave simulations have a periodic structure in substantial portions of the signal that noise-band simulations do not have. An algorithm applied in acoustic periodicity detection on the basis of a noise-resistant autocorrelation method (Boersma, 1993) was capable of detecting consecutive pitch periods in an important proportion of the processed sine-wave signal. As expected, the same method failed to find pitch periods in the noise-band simulations. We hypothesize that one advantage of sine-wave versus noise carriers is in the time-amplitude envelope. Noise carriers have a rapidly fluctuating envelope whereas the sine wave has a fixed amplitude envelope. When the speech envelope is extracted and imposed on one of these carriers, random level variations in the noise carrier serves to distort the speech envelope. Noise-band simulations only provide information on the spectral distribution of energy. With this type of stimuli, listeners should base their perceptual judgments mainly on the rough spectral information carried by noise bands. In this case, the frequency resolution of the signal would increase as the processor implements more channels and this would explain why performance improved substantially with the number of channels.

A plausible interpretation for the observed results is in the modulation domain. Using a noise carrier, normal hearing subjects can detect typically 5%–10% amplitude modulation (Viemeister, 1979), whereas they can detect 1%–5% modulation when a sinusoidal carrier is used (Kohlrausch *et al.* 2000). This latter observation is more in line with the cochlear-implant user's ability to detect amplitude modulation (Shannon, 1992). The sine-wave carrier is a single frequency component, and when the speech envelope is imposed on this kind of carrier, side bands are generated reflecting the spectral content of the envelope. In the case of the noise carrier, these side bands are masked because they coincide with spectral components of the noise. The better result with the sine-wave simulations is likely due to (a) better modulation detection with the sine wave than with the noise carrier, and (b) the likely resolved side bands of amplitude modulation, particularly in the low-frequency bands. The modulation introduces spectral side bands, which may be detected as separate components if they are sufficiently far in frequency from the carrier frequency (Kohlrausch *et al.* 2000). In this sense, the better performance with sine-wave carriers would be due to a better representation of modulation as well as to the perception of the resolved side bands. From this point of view, the sine-wave stimuli could simulate the implant performance better than noise-band stimuli, supporting Zeng's assertion made by making quantitative comparisons between performance of normal and implant listeners (Zeng, 2004). On the other hand, one possibility is that some Fo information from the original signal remains in the sine-wave stimulus. The 160 Hz cutoff frequency with a second-order low-pass filter may still be too high to prevent leakage of Fo information into the envelope domain.

Roughly speaking, performance in gender and speaker identification increases as the number of channels increases, but this effect is more obvious in the noise band processor. Sine-wave stimuli attain a very high performance even at the fewest number of channels, giving rise to a ceiling effect that prevents a clear channel effect. Only in the most difficult task used for speaker recognition, i.e., across different utterances, the channel effect is significant for the sine-wave processor.

On the other hand, comparing results from Experiments 2 and 3 it is clear that speaker identification is easier across the same utterance than through different utterances. Linguistic and probably other idiosyncratic acoustic cues have contributed to the difference between both experiments. This is congruent with recent research by McDonald *et al.* (2003), where differences in talker discrimination by CI and normal-hearing listeners was studied under two linguistic conditions. In the first one, listeners heard pairs of stimuli (words or sentences) whose linguistic content was identical (e.g., cat–cat). In the second condition, the linguistic content of each pair was different (e.g., cat–dog). Discrimination accuracy was better in stimulus pairs where the linguistic content was held constant. In the same study it was found that talker discrimination was easier for male–female talker pairs than for within-gender stimulus pairs. Partial analysis of our data from Experiments 2 and 3 by separating trials between same-gender versus different-gender subsets yielded the same pattern of results. At the same time, our data clearly showed that speaker identification occurred beyond the recognition of gender, i.e., in the same-gender subset.

Replication of Experiments 1–3 using a 160 Hz cutoff frequency showed the same results as with the 400 Hz cutoff

frequency, except for the most difficult condition, in which there is less spectral information (3 channels) for speaker recognition. Results presented by Chinchilla and Fu (2003b) fit well with our data. These authors studied voice gender discrimination by CI and normal-hearing (NH) listeners using sine-wave and noise-band vocoders simulations, where number of channels (4 to 32) and the cutoff frequencies of the channel's envelope filters (from 20 to 320 Hz) were manipulated. Results for NH subjects with sine-wave stimuli showed that when only 4 spectral channels were available, gender discrimination improved as the envelope filter cutoff frequency was increased from 20 to 320 Hz. In other words, both spectral and temporal information contribute to gender and speaker identification, but the temporal information is especially important when there are few spectral cues for the identification task.

A last issue to be considered is which of the two processors is a better approximation to the actual performance of cochlear-implant users. When the focus of research is on speech intelligibility, the small differences found between both processors do not allow a clear conclusion on the matter. Empirical data about speech intelligibility seem to indicate that both processors work reasonably well as simulations of the CI signal (Dorman *et al.*, 1998; Fu *et al.*, 1998; Loizou *et al.*, 2000; Friesen *et al.*, 2001; Loizou and Poroy, 2001). However, results on gender and speaker recognition are very different depending on the nature of the output signal. To date, there are not enough data to support a conclusive statement. Chinchilla and Fu (2003b) found that gender discrimination scores were highly variable among CI listeners. However, the best-performing CI users scored similarly to normal-hearing (NH) subjects listening to the 4-channel sine-wave simulations. Based on this similarity, the authors suggested that sine-wave vocoder simulations may better approximate CI user's listening conditions than noise-band processors. Zeng (2004) considers that sine-wave stimuli can simulate CI performance better than noise-band stimuli because the CI listener's ability to detect amplitude modulation is more similar to the NH listener's ability when using a sinusoidal carrier than when a noise carrier is used. In any case, we think it is too soon for drawing firm conclusions. In speaker recognition there is a lack of data comparing CI's and NH's performance through sine-wave and noise-band processors. However, Chinchilla and Fu's data on voice gender discrimination show a good fit between the *best-performing* CI users and NH subjects listening sinewave stimuli processed through the *fewest* number of channels. In this respect, further studies on performance of non-best-performing CI users would be valuable, since they would perhaps show a good fit to the noise-band processor, which yields lower performance. Further research is needed to compare gender/speaker recognition performance between CI users and NH subjects listening to sine-wave and noise-band simulations.

ACKNOWLEDGMENTS

This study was partly supported by *Fundació Caixa Castelló-Bancaixa* and the Universitat Jaume I, Castellon (Project No. P1.1A2002-01) and the Ministry of Science and Technology of Spain (I+D+I, Project No. BSO2003-01002/PSCE). The authors would like to thank Philipos C. Loizou, who kindly provided by email his MATLAB routines to create the simulations based on the sine-wave processor. We would also like to thank Paul Boersma and David Weenink for their PRAAT software and Chris Darwin for his PRAAT script, since a modified version of it was used to create the simulations based on the noise-band processor. The authors would like to thank the Associate Editor, Dr. Kenneth W. Grant and Dr. Fan-Gang Zeng for helpful and valuable comments received on an earlier version of this paper. Most of the explanations for the superiority of the sine waves were suggested by Dr. Grant and Dr. Zeng.

[1]Signal processing with the sine-wave processor was implemented using the MATLAB routines (*Csim.m, Estfilt.m, Gethdr.m*, and *Mel.m*), kindly provided by Philipos C. Loizou (Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX).

[2]Signal processing with the noise-band processor was implemented by means of the PRAAT software (Boersma and Weenink, 2001) using a modified version of a script provided by Chris Darwin (Laboratory of Experimental Psychology, University of Sussex, Brighton, UK) at the web address: http://www.biols.susx.ac.uk/home/Chris_Darwin/Praatscripts/Shannon.

Boersma, P. (**1993**). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," Proc. Institute of Phonetic Sciences **17**, 97–110.

Boersma, P., and Weenink, D. (2001). *Praat 4.0: A system for doing phonetics by computer* (computer software) (University of Amsterdam, Amsterdam, The Netherlands). Available online: http://www.praat.org

Bricker, P. D., and Pruzansky, S. (**1976**). "Speaker recognition," in *Contemporary Issues in Experimental Phonetics*, edited by N. J. Lass (Academic, New York), pp. 295–326.

Chinchilla, S. S., and Fu, Q. J. (2003a). "Discrimination and vowel recognition in normal-hearing and cochlear implant users," in Abstracts of the 26th Midwinter Research Meeting of the Association for Research in Otolaryngology.

Chinchilla, S. S., and Fu, Q. J. (2003b). "Voice gender discrimination and vowel recognition in normal-hearing and cochlear implant users," in Abstracts of the Conference on Implantable Auditory Prostheses, Asilomar, CA.

Cleary, M., and Pisoni, D. (**2002**). "Talker discrimination by prelingually deaf children with cochlear implants: Preliminary results," Ann. Otol. Rhinol. Laryngol. Suppl. **189**, 113–118.

Dorman, M., Loizou, P., 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.

Dorman, M., Loizou, P., and Rainey, D. (**1997**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411.

Fellowes, J., Remez, R., and Rubin, P. (**1997**). "Perceiving the sex and identity of a talker without natural vocal timbre," Percept. Psychophys. **59**, 839–849.

Friesen, L., Shannon, R. V., Baskent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163.

Fu, Q-J., Shannon, R., and Wang, X. (**1998**). "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," J. Acoust. Soc. Am. **104**, 3586–3596.

Fujita, S., and Ito, J. (**1999**). "Ability of nucleus cochlear implantees to recognize music," Ann. Otol. Rhinol. Laryngol. **108**, 634–640.

Gfeller, K., Witt, S., Woodworth, G., Mehr, M. A., and Knutson, J. (**2002**). "Effects of frequency, instrumental family, and cochlear implant type on

timbre recognition and appraisal," Ann. Otol. Rhinol. Laryngol. **111**, 349–356.
Gfeller, K., Woodworth, G., Robin, D. A., Witt, S. G., and Knutson J. (**1997**). "Perception of rhythmic and sequential pitch patterns by normally hearing adults and cochlear implant users," Ear Hear. **18**, 252–260.
Hiki, S., and Fukuda, Y. (**2000**). "Pitch perception through the cochlear implant for speech and music," Adv. Oto-Rhino-Laryngol. **57**, 12–24.
Hoesel, R. J. M., Ramsden, R., and Odriscoll, M. (**2002**). "Sound-direction identification, interaural time delay discrimination, and speech intelligibility advantages in noise for a bilateral cochlear implant user," J. Acoust. Soc. Am. **23**, 137–149.
Hoesel, R. J. M., and Tyler, R. S. (**2003**). "Speech perception, localization, and lateralization with bilateral cochlear implants," J. Acoust. Soc. Am. **113**, 1617–1630.
Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.
Kohlrausch, A., Fassel, R., and Dau, T. (**2000**). "The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers," J. Acoust. Soc. Am. **108**, 723–734.
Kong, Y.-Y., Vongphoe, M., and Zeng, F.-G. (2003). "Independent contributions of amplitude and frequency modulations to auditory perception. II. Melody, tone, and speaker identification," in Abstracts of the 26th Midwinter Research Meeting of the Association for Research in Otolaryngology.
Kreiman, J. (**1997**). "Listening to voices: Theory and practice in voice perception research," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix (Academic, San Diego, CA), pp. 85–108.
Lass N. J., Almerino, C. A., Jordan, L. F., and Wals, J. M. (**1980**). "The effect of filtered speech on speaker race and sex identification," J. Phonetics **8**, 101–112.
Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., and Bourne, V. T. (**1976**). "Speaker sex identification from voiced, whispered and filtered isolated vowels," J. Acoust. Soc. Am. **59**, 675–678.
Lobo, A. P., Toledos, F., Loizou, P., and Dorman, M. F. (**2002**). "Effect of envelope lowpass filtering on consonant melody recognition," J. Acoust. Soc. Am. **112**, 2245.
Loizou, P., Dorman, M., Poroy, O., and Spahr, T. (**2000**). "Speech recognition by normal-hearing and cochlear implant listeners as a function of intensity resolution," J. Acoust. Soc. Am. **108**, 2377–2387.
Loizou, P., Dorman, M., and Tu, Z. (**1999**). "On the number of channels needed to understand speech," J. Acoust. Soc. Am. **106**, 2097–2103.
Loizou, P., and Poroy, O. (**2001**). "Minimal spectral contrast needed for vowel identification by normal hearing and cochlear implant listeners," J. Acoust. Soc. Am. **110**, 1619–1627.
McDermott, H. J., and McKay, C. M. (**1997**). "Musical pitch perception with electrical stimulation of the cochlea," J. Acoust. Soc. Am. **101**, 1622–1631.
McDonald, C. J., Kirk, K. I., Krueger, T., and Houston, D. (2003). "Talker discrimination by adults with cochlear implants," in Abstracts of the 26th Midwinter Research Meeting of the Association for Research in Otolaryngology.
Mullennix, J. W., Johnson, K. A., Topcu-Durgun, M., and Farnsworth, L. M. (**1995**). "The perceptual representation of voice gender," J. Acoust. Soc. Am. **98**, 3080–3095.
Qin, M., and Oxeham, A. (**2003**). "The effects of simulated cochlear-implant processing on F0 discrimination," J. Acoust. Soc. Am. **113**, 2224.
Remez, R., Fellowes, J., and Rubin, P. (**1997**). "Talker identification based on phonetic information," J. Exp. Psychol. Hum. Percept. Perform. **23**, 651–666.
Shafiro, V., Jenkins, J., and Strange, W. (**2003**). "Identifying the sources of environmental sound with a varying number of spectral channels," J. Acoust. Soc. Am. **113**, 2326.
Shannon, R., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Shannon, R., Zeng, F.-G., and Wygonski, J. (**1998**). "Speech recognition with altered spectral distribution of envelope cues," J. Acoust. Soc. Am. **104**, 2467–2476.
Shannon, R. V. (**1992**). "Temporal modulation transfer functions in patients with cochlear implants," J. Acoust. Soc. Am. **91**, 2156–2164.
Sheffert, S. M., Pisoni, D. B., Fellowes, J. M. and Remez, R. E. (**2002**). "Learning to recognize talkers from natural, sinewave, and reversed speech samples," J. Exp. Psychol. Hum. Percept. Perform. **28**, 1447–1469.
Spahr, A., Dorman, M., and Loizou, P. (**2002**). "Effects on performance of partial misalignments of spectral information in acoustic simulations of cochlear implants," J. Acoust. Soc. Am. **112**, 2356.
Viemeister, N. F. (**1979**). "Temporal modulation transfer function based on modulation thresholds," J. Acoust. Soc. Am. **66**, 1364–1380.
Wu, K., and Childers, D. G. (**1991**). "Gender recognition from speech. I. Coarse analysis," J. Acoust. Soc. Am. **112**, 1828–1848.
Zeng, F. G. (**2004**). "Compression and cochlear implants," in *Springer Handbook of Auditory Research; Compression: From Cochlea to Cochlear Implants*, edited by S. P. Bacon, R. R. Fay, and A. N. Popper (Springer, New York), pp. 184–220.