Voice gender perception by cochlear implantees

Kovačić, Damir; Balaban, Evan

Voice gender perception by cochlear implantees[a)]

Damir Kovačić[b)]
Cognitive Neuroscience Sector, Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Lionello Stock 2/2, 34135 Trieste, Italy and SUVAG Polyclinic, Kneza LJ. Posavskog 10, 10000 Zagreb, Croatia

Evan Balaban
Behavioral Neurosciences Program, Stewart Biological Sciences Building, McGill University, 1205 Avenue Dr. Penfield, Montreal, Quebec H3A 1B1, Canada and Laboratorio de Imagen Médica, Hospital General Universitario Gregorio Marañón, Calle Dr. Esquerdo 46, 28007 Madrid, Spain

(Received 5 June 2008; revised 4 March 2009; accepted 26 May 2009)

Gender identification of human voices was studied in a juvenile population of cochlear implant (CI) users exposed to naturalistic speech stimuli from 20 male and 20 female speakers using two different voice gender perception tasks. Stimulus output patterns were recorded from each individual CI for each stimulus, and features related to voice fundamental frequency and spectral envelope were extracted from these electrical output signals to evaluate the relationship between implant output and behavioral performance. In spite of the fact that temporal and place cues of similar quality were produced by all CI devices, only about half of the subjects were able to label male and female voices correctly. Participants showed evidence of using available temporal cues, but showed no evidence of using place cues. The implants produced a consistent and novel cue to voice gender that participants did not appear to utilize. A subgroup of participants could discriminate male and female voices when two contrasting voices were presented in succession, but were unable to identify gender when voices were singly presented. It is suggested that the nature of long-term auditory categorical memories needs to be studied in more detail in these individuals.
 [DOI: 10.1121/1.3158855]

PACS number(s): 43.64.Me, 43.71.Bp, 43.66.Ts [RYL] Pages: 762–775

I. INTRODUCTION

Cochlear implant (CI) users exhibit high variability in their ability to perceive speech (Svirsky *et al.*, 2000; Blamey *et al.*, 2001). An important goal of CI research is to elucidate potential factors underlying this variation. Anecdotal evidence and previous studies (Cleary and Pisoni, 2002; Fu *et al.*, 2004; Spahr and Dorman, 2004; Cleary *et al.*, 2005; Fu *et al.*, 2005) suggest that even seemingly simple tasks such as identifying speakers pose challenges for CI users.

The identification of voice gender must be ultimately based on the quality of the spectral and temporal cues that speech items provide, even though individual variation in each individual's memory, experience, and motivation are also necessarily involved. Gender-related features of vocal tract anatomy [via body size effects on vocal tract length (VTL)], and laryngeal fold size [influencing differences in the temporal rate of vibration, which are reflected in differences in fundamental frequency (F0)] provide normal-hearing (NH) listeners with cues for voice gender identification. An almost perfect recognition of voice gender is achieved in NH individuals when both VTL and F0 cues are used (Bachorowski and Owren, 1999; Owren *et al.*, 2007; Smith *et al.*, 2007). In addition, automated recognition systems using VTL and F0 show robust performance in the face of within- and between-speaker acoustic variations (Childers and Wu, 1991; Wu and Childers, 1991).

CIs appear to elicit temporal pitch percepts only below ~300 Hz (Carlyon and Deeks, 2002; Zeng, 2002) and have inherently low spectral resolution (Cohen *et al.*, 1996; McKay, 2005). CI users may rely on particular isolated spectral or temporal cues, or some combination in gender identification tasks (Fu *et al.*, 2004; Laneau and Wouters, 2004b; Fu *et al.*, 2005; Chang and Fu, 2006). Temporal cues are delivered in the form of envelope modulations to one or more stimulation electrodes, while spectral cues are provided via the spatial pattern of electrode array stimulation (McKay *et al.*, 1994, 1995; Laneau and Wouters, 2004a, 2004b). Such cues should yield high levels of gender identification when voices are far apart from each other in F0 and VTL, even though CI users are poor at distinguishing speakers with similar vocal characteristics (Cleary and Pisoni, 2002; Fu *et al.*, 2005). However, the extent to which the implants of different users provide information of similar quality about natural speech signals may also play a role in individual variation in CI listener performance.

The present study examines CI users' perception of voice gender by assessing the degree to which individual CIs reliably transmit potentially available voice gender cues to their wearers, and the dependence of behavioral performance on the quality of this information.

[a)]Portions of these data were presented at the Conference for Implantable Auditory Prosthesis 2007, Lake Tahoe, CA.

[b)]Author to whom correspondence should be addressed. Present address: Laboratory of Auditory Neurophysiology, K.U. Leuven, Campus Gasthuisberg O&N 2, Herestraat 49 bus 1021, B-3000 Leuven, Belgium. Electronic mail: Damir.Kovacic@med.kuleuven.be

A. Availability of temporal cues to voice gender

Previous studies examining signals with amplitude modulated envelopes, both in hearing subjects (Burns and Viemeister, 1976, 1981) and CI users (McKay *et al.*, 1995; McKay and McDermott, 2000), show that the fundamental frequency of the input signal can be effectively encoded by temporal modulations of the signal envelope. Within the signal processor of each implant, input signals are processed into frequency sub-bands by means of a set of bandpass filters. Each sub-band is then assigned to the most relevant electrode in such a way that tonotopic organization is roughly preserved. Sub-bands with high-frequency content are delivered to the most basal electrodes, and sub-bands with low-frequency information are delivered to electrodes closer to the cochlear apex. In order to minimize perception of the carrier frequency and to correctly represent envelope modulations related to F0, the carrier frequency is usually at least four times higher than the typical F0 value (McDermott and McKay, 1994).

To find channels that are likely to represent F0 in their envelopes, the modulation spectrum (MS) of the stimulus sounds was examined after being processed by the same filter-bank settings used in each subject's device. By comparing the acoustic MS (calculated from bandpass-filtered acoustic signals) with MSs obtained from electrodograms captured from a subject's device, the electrode channel(s) that are theoretically able to deliver F0-related temporal cues relevant for gender identification in each subject could be identified. These channels are said to have F0 modulation availability, and were used to ascertain the relationship between perceptual performance and the quality of available F0 information for the different stimuli.

B. Availability of place cues to voice gender

Previous research with both CI devices and CI simulations delivered to NH subjects suggests that place cues for voice gender identification are available to CI users (Fu *et al.*, 2004, 2005; Gonzalez and Oliver, 2005). Laneau and Wouters (2004a) also demonstrated possible utilization of place cues for pitch discrimination in multi-electrode stimulation settings, with the average just-noticeable differences for place pitch ranging from 0.25 to 0.46 mm.

The present study assessed the availability of place cues for voice gender based on the relationship between stimulated electrode location and the vocal pitch or spectral envelope distribution of the speakers. Each stimulus produced a specific spatial output profile of electrode stimulation, quantified here as the cumulative sum of pulses in each electrode during the 2 s of stimulation provided by each stimulus. A median central location for this electrode stimulation pattern, equivalent to the "center-of-mass" of the area below the curve in the stimulation output profile, was calculated. The correlation between median electrode values and either the acoustic F0 of speech items or the distribution of their spectral envelope energy was then examined.[1]

C. Overall design of the experiments

Quantitative measures of temporal and place information that the implants provide about vocal stimuli were related to variation in behavioral responses to these same stimuli using two different testing procedures. The first was a fixed identification procedure with a one-interval, two-alternative forced choice (2AFC) task in which subjects indicated whether the speech sample was spoken by a male or female speaker. The second was an adaptive two-interval 2AFC discrimination procedure in which the adaptive parameter was the difference in the fundamental frequencies (F0) of male and female speech items (subjects responded by indicating which speech item was uttered by a female speaker). These two procedures were employed in order to assess the potential role of long-term auditory representations in CI users' performance. Allowing direct, short-term comparisons of two speech samples permitted additional evaluation of the auditory abilities of subjects who were not able to distinguish voice gender in the fixed procedure. The adaptive procedure also directly assessed how voice gender identification abilities relate to the size of F0 differences. The two procedures employed different speech items spoken by the same set of 20 male and 20 female speakers. This number of speakers was used to ensure that general processing strategies for the differences between male and female voices were being studied, rather than specific memorization strategies that could be used to discriminate among a small number of voices.

II. METHODS

A. Subjects

The study was approved by the ethical committees of the School of Medicine, University of Zagreb, Polyclinic SUVAG, and the Croatian Medical Chamber. Forty-one CI subjects with devices manufactured by Cochlear Corporation (20 males and 21 females; age range: 5.3–18.8 years, mean age=12.3 years) were recruited into the study using a database maintained by the Polyclinic SUVAG (a Croatian national institute for the rehabilitation of listening and speech). Subject details are given in Table I. One purpose of this research was to establish the relationship between the signal information being delivered to each individual CI user and their perceptual behavior. Since the apparatus for capturing stimulus output patterns was only provided by one CI manufacturer (Cochlear Ltd.), and was not compatible with devices of other manufacturers, it was necessary to limit the subject pool to individuals using Cochlear Ltd. devices (Esprit 3G and Sprint).

Psychological assessments of the participants were consulted to disqualify any subjects with reduced cognitive abilities that might influence their gender identification performance (Waltzman *et al.*, 2000; Holt and Kirk, 2005; Stacey *et al.*, 2006). Nonverbal psychological assessment included at least two or more of the following tests: Raven progressive matrices, Goodenough IQ, nonverbal WISC IQ, Brunet–Lezine, and Leiter International Performance Scale. These tests were performed by professional psychologists at least once during the subject's association with the Polyclinic

TABLE I. Characteristics of the CI participant population. Asterisks (*) denote unsuccessful capture of CI electrical signals.

Subject	Age at testing (years; months)	Sex	Ear	Processor type	No. of electrodes	Stimulation rate per channel (Hz)	Sound coding strategy
CI01	5; 4	F	R	Esprit 3G	20	1200	ACE
CI02	10; 6	M	R	Sprint	22	1200	ACE
CI03	12; 1	M	L	Esprit 3G	20	900	ACE
CI04	15; 8	F	R	Esprit 3G	20	900	ACE
CI05	18; 9	M	L	Esprit 3G	20	900	ACE
CI06	11; 4	M	R	Sprint	21	1200	ACE
CI07	12; 2	M	R	Esprit 3G	20	900	ACE
CI08	6; 9	M	L	Esprit 3G	20	1200	ACE
CI09	9; 5	M	R	Esprit 3G	20	1200	ACE
CI10	11; 0	F	L	Esprit 3G	20	900	ACE
CI11	11; 1	M	R	Esprit 3G	20	900	ACE
CI12	14; 0	M	R	Esprit 3G	20	900	ACE
CI13*	14; 7	F	R	Esprit 3G	20	250	SPEAK
CI14	11; 3	M	L	Esprit 3G	20	900	ACE
CI15	17; 6	F	R	Esprit 3G	20	900	ACE
CI16	7; 9	M	R	Esprit 3G	20	1200	ACE
CI17	14; 5	F	R	Esprit 3G	20	1200	ACE
CI18	13; 1	F	R	Esprit 3G	20	900	ACE
CI19*	14; 4	F	R	Esprit 3G	18	500	ACE
CI20*	9; 7	F	R	Esprit 3G	20	900	ACE
CI21	11; 0	M	R	Sprint	19	1200	ACE
CI22	18; 5	F	L	Esprit 3G	20	900	ACE
CI23	9; 2	F	R	Esprit 3G	20	1200	ACE
CI24	14; 9	F	R	Esprit 3G	20	900	ACE
CI25	8; 8	F	R	Esprit 3G	19	1200	ACE
CI26	17; 3	M	R	Esprit 3G	20	900	ACE
CI27*	12; 9	F	R	Esprit 3G	16	900	ACE
CI28*	8; 5	F	L	Esprit 3G	19	1200	ACE
CI29	11; 4	M	R	Esprit 3G	20	1200	ACE
CI30	8; 9	M	R	Sprint	22	1200	ACE
CI31	15; 7	M	L	Esprit 3G	20	1200	ACE
CI32	13; 11	F	R	Esprit 3G	20	900	ACE
CI33	14; 0	F	R	Esprit 3G	20	900	ACE
CI34	12; 11	F	R	Esprit 3G	20	900	ACE
CI35*	14; 7	F	R	Esprit 3G	19	720	ACE
CI36	10; 5	F	R	Esprit 3G	20	1200	ACE
CI37	15; 10	M	R	Esprit 3G	20	900	ACE
CI38	10; 11	M	R	Esprit 3G	19	1200	ACE
CI39	9; 1	M	R	Sprint	22	1200	ACE
CI40	14; 4	M	L	Esprit 3G	20	900	ACE
CI41	10; 0	F	L	Sprint	22	900	ACE
Mean	12.3						
SD	3.2						

SUVAG as a part of their hearing rehabilitation care. None of the subjects who participated in the present study had significant nonverbal psychological disorders. Spoken language proficiency was compromised in some cases because of previous auditory deprivation (Svirsky *et al.*, 2000, 2004). However, no minimal requirements in language proficiency were used as part of the selection criteria for the present study. Both the simplicity of the gender identification task (not requiring complex linguistic skills), and the fact that every participant in the study was able to provide their own explanation of what they were supposed to do, suggest that variations in spoken language proficiency did not create problems for subjects understanding the nature of the experimental tasks.

Control participants were 15 hearing children (8 males and 7 females, age range 6.7–10.6 years, mean age 9.3 years) recruited from one primary school in Zagreb. Neither the participants and their parents, nor their teachers reported hearing problems. In each case, parents or caregivers signed a consent form before their children participated in this study. The sex composition did not differ between the control (C) and experimental (E) group ($\chi^2=0.09$, $p>0.99$), but

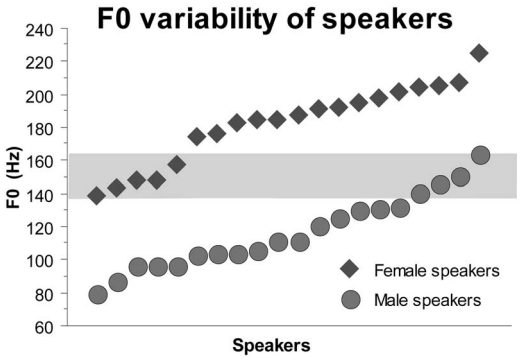

FIG. 1. Fundamental frequency (F0) values for the 20 male and 20 female speakers used in this study. The mean female F0 = 183.3 ± 5.4 Hz, and the mean male F0 = 117.9 ± 4.8 Hz. The shaded horizontal band denotes the region of overlapping fundamental frequencies (between 137.9 and 163.3 Hz). With overlapping speakers excluded, the mean female F0 = 195.1 ± 3.1 Hz and the mean male F0 = 110.1 ± 3.3 Hz.

the age composition differed (Mann–Whitney U-test, $\text{mean}_C = 9.3$, [SD = 1.3] years, $\text{mean}_E = 12.3$, [SD = 3.2] years, $n_C = 15$, $n_E = 41$, $p < 0.001$). The primary goal of the control group was to check whether the stimuli were of abnormal difficulty; due to the reduced hearing experience of CI subjects and their delays in language acquisition (Svirsky *et al.*, 2000, 2004; Nicholas and Geers, 2007), it is appropriate to compare older CI subjects with younger hearing controls.

B. Stimuli

Speech samples in the form of short news-like stories from 20 different male and 20 different female professional radio announcers were obtained from the national broadcasting radio company *Hrvatski radio* in digitized format, sampled at 44100 Hz using a 16-bit coding scheme. The samples were cut into 40 speech items of 2-s length with the following two requirements: (i) the onset was always aligned with the word onset and (ii) the offsets were never within phoneme boundaries. Since it was not possible to maintain the exact length of 2000 ms for all speech items in this way, the PSOLA lengthening algorithm (Moulines and Laroche, 1995) was used to equalize the length of all utterances. This manipulation was done using PRAAT software (Boersma and Weenink, 2006) with scaling factors between 0.84 and 1.25. At these scaling values, the PSOLA lengthening algorithm preserves pitch contours. Fundamental frequencies of the speakers were calculated using the autocorrelation method (Boersma, 1993) implemented in PRAAT software. The average F0 values of the speech items are shown in Fig. 1. Natural F0 variability occurs in this stimulus set both within the utterances of each speaker (pitch contours not shown) and between speakers. The male and female stimulus populations differed in their average fundamental frequencies [average female F0 = 183.3 ± 5.4 Hz and average male F0 = 117.9 ± 4.8 Hz, about 56% of one octave apart ($Z = 5.13$, $n = 20$, $p < 0.0001$, Mann–Whitney U-test)]. Natural variability in this sample yielded an overlapping area of F0 values between five female and four male voices, as indicated in Fig. 1 by the horizontal band. This overlapping range was between 137.9 (minimal female F0) and 163.3 Hz (maximal male F0). With overlapping speakers removed, the mean female F0 was 195.1 ± 3.1 Hz and the mean male F0 was 110.1 ± 3.3 Hz.

C. Voice gender identification

Prior to the commencement of the experiment, each subject was given instructions and performed practice trials until the experimenter was assured that the subject understood the task. The practice and experimental trials used PRESENTATION software (Version 10.1, Neurobehavioral Systems, Inc., Albany, CA). In each trial, one 2-s-long speech item was chosen randomly, and delivered to either headphones (hearing control subjects) or to the CI device (CI subjects) via direct line input. The subject was requested to respond by clicking on one of two response buttons representing a male and a female. The response buttons were associated with a computer display showing sketches of typical male and female faces. Feedback (a smiling face for correct responses and the symbol "X" for incorrect responses) was given in order to maintain the subject's interest in the experiment. The items were not replayed, even for incorrect responses. The first and second halves of the trials had response button positions exchanged (in the first half, the female was associated with the left button). Five sets of randomly chosen speech items were created and their presentation was counterbalanced between subjects. Practice consisted of six trials with two different female and male speech items from the same speaker database that were not used in the experimental trials. Response buttons (indicating male or female voice) were interchanged after half of the practice trials. CI subjects typically needed just one practice run before understanding the task and starting the experimental phase. To make sure that the participant understood the task, they were instructed to give their own explanation of what they had to do to the experimenter. Since the aim of this experiment was to assess each participant's ability to identify the gender of a voice, their performance did not have to reach a pre-defined criterion in order to progress to the experimental phase. Both the simplicity of the task, and the visual reinforcement with sketched pictures of male and female faces during response periods allowed all subjects (including two 4-year-old hearing subjects in a pilot study) to perform the task reliably throughout the practice trials and the experiment.

For the hearing control subjects, Sennheiser HD 580 headphones were used, with monaural presentation of stimuli to the right ear at 65 dB sound pressure level (SPL) (*A*-level) as measured by a RadioShack sound level meter (model 33-2055). Except for this difference in the stimulus delivery apparatus, the items and procedures were the same in both hearing control and CI subjects. After completion of the fixed identification procedure, the subjects had a short break (10 min) and then resumed their participation by undergoing a second procedure, described below. When the second adaptive procedure finished, the electrode stimulation patterns generated by all of the presented speech items were captured using processor control interface (PCI) and interface card (IF5) hardware (manufactured by Cochlear Ltd., Sydney, Australia and provided by Cochlear AG, Basel, Switzerland)

and the RFSTAT software system [manufactured and provided by the Cooperative Research Centre for Cochlear Implant and Hearing Aid Innovation (CRC HEAR), Melbourne, Australia]. CI signal capture was performed by recording two long audio sequences played to the subject's CI device with the transmitting coil attached to the PCI+IF5 system (instead of the subject's implanted receiver coil), keeping the same device settings used during the experiment. Each audio sequence was 84.25 s long and contained all speech items presented in the two procedures. Speech items were sequentially placed one after the other in the audio sequence and were separated by 100 ms of silence. In order to synchronize the capture signal onset with that of the audio signal, a 50-ms white noise burst was placed at the beginning of each long sequence. The analysis of CI captured signals was performed using custom software written in MATLAB (The Mathworks, Natick, MA). While the signals were being captured, subjects were led to an adjoining room and asked to make drawings of familiar objects. Total time of the experiment was typically about 30–40 min (fixed and adaptive procedures typically lasting about 5–10 min each).

D. Adaptive speech-based F0 discrimination of voice gender

An adaptive speech-based F0 discrimination procedure was used to assess voice gender perception as a function of voice F0 differences using an up-down staircase method (Levitt, 1971). Since several researchers have highlighted the susceptibility of adaptive threshold estimates to variability caused by attentional lapses or confusion at the beginning of trials (Baker and Rosen, 2001; Amitay *et al.*, 2006), subjects were first instructed how to respond in this task. They were told that they were going to hear two speech excerpts, one after the other, in which one item was always male and the other was always female, with no possibility that both were male or female. Their task was to listen to both speech items and choose by pressing the appropriate button on the keyboard which one was female. The female/male distinction was reinforced through analogy with the child's mother and father, and, a practice test was administered with ten trials. The majority of CI subjects successfully completed the practice test confirming that they understood the task; the others were all successful after repeating the practice test once more.

The stimuli were recordings of different utterances from the same speakers used in the identification procedure (no utterances were shared between procedures). In each trial, one speech item selected from 20 speakers of one sex was paired with another speech item selected from 20 speakers of the other sex. The position of the female voice in the stimulus pair was chosen at random. The stimulus set consisted of 400 different stimulus pairs, each of which had a unique ΔF0 between the female and male speech item. The largest separation of F0 in a stimulus pair was 142.8 Hz, and there was a region with negative ΔF0 for stimulus pairs in which the F0 of the female voice was lower than the F0 of the male voice (with the minimal value of $\Delta F0 = -23.6$ Hz).

The initial ΔF0 was set to 109 Hz, and the adaptive step was held constant at 10 Hz. The experiment stopped when either 10 reversal points had been achieved or 50 total trials had been run. In order to avoid possible pop-out memory effects of repeating stimulus pairs that were played recently, a minimal span of four trials was set between repeats of the same stimulus pair. In accordance with the relative scarcity of stimulus pairs on both extreme ends of the distribution of ΔF0, ΔF0 could not be increased or decreased beyond these edge points. To deal with cases in which subjects would become stuck at the edges, the adaptive procedure was modified; if the ΔF0 entered the region consisting of the five stimulus pairs with either the largest or the smallest F0 differences, then the adaptive procedure would randomly choose one of the stimulus pairs in this range.

Performance in the adaptive discrimination procedure was assessed with discrimination threshold estimates (DTEs). These were obtained for each subject by averaging the values of the last five reversal points of the adaptive procedure. Since the value of the adaptive parameter was limited by natural variation in the population of speakers, the subject might reach either a ceiling (indicating perfect performance) or floor level (indicating chance performance). In these cases, DTEs cannot be estimated more precisely than being smaller or larger than the edge boundaries. The lower boundary in this procedure was $\Delta F0 = -13.5$ Hz, whereas the upper boundary was $\Delta F0 = 133.1$ Hz.

Hearing controls were run using the same stimuli and procedures as implemented for the CI group, the only difference being the presentation of the sounds through Sennheiser HD 580 headphones [monaural presentation to the right ear at 65 dB SPL(*A*)].

E. Capturing procedures of stimulus-induced electrode output patterns from CI devices

In order to make sure that the CIs received direct stimulation from the experimental computer, without being contaminated by external background noise, a specially-constructed sound-insulated chamber was used (see Appendix). To maintain comparable loudness levels corresponding to 65 dB SPL(*A*) for all stimulus presentations to each subject, a two-step procedure was followed. The level of each stimulus was first calibrated by manual manipulation of signal levels in order to have a RadioShack model 332055 sound level meter reading of 65 dB SPL(*A*) in free-field, approximately 105 cm from the midline of a Harman Kardon 2.1 loudspeaker system. Then, using each subject's clinically-assigned CI processor in isolation from the subject, the master volume level on the computer playback software was adjusted by playing a standard 10-s multi-talker babble stimulus free-field via a Soundblaster Creative Audigy ZS sound card and Harman Kardon 2.1 loudspeaker system. The sound level meter was located at the Esprit 3G or Sprint microphone approximately 120 cm from the midline of the loudspeaker, and was used to measure the intensity level while adjusting the master/output level of the signal to produce a sound pressure level of 65 dB SPL(*A*). Once the proper setting was achieved, RF STATISTICS software was used to capture the CI output while playing each stimulus through the CI processor. Using the "Statistics" tab of the

capture software, it was possible to determine the average current level over the duration of each stimulus. The average of these current levels was taken, and the 10-s multi-talker babble stimulus was then played to the CI processor while the stimulus level was adjusted to produce an average current level identical to that obtained for the experimental stimuli. The CI processor was then returned to the subject, and this final setting was used to play the experimental stimuli to them. The relationship between the clinical units measured by the capturing system and actual physical (current) units (I in μA) was determined by the following equation:

$$I = F \times 10 \times 175^{\mathrm{cl}/255}, \quad (1)$$

where F is the calibration factor (equal to 1 for the devices studied here, Peter Seligman, Cochlear Ltd., personal communication) and cl is the value in clinical units (Drennan and Pfingst, 2006).

F. Analysis procedures for captured data

The 1 μs temporal resolution of the capturing apparatus was changed to 20 μs (corresponding to a sampling frequency of 50 kHz) using time-scale conversion, in order to decrease the memory burden for computational analyses. Each channel in each subject's device had minimum (threshold or T-level) and maximum (comfortable or C-level) current level values, which bracketed the device's stimulation levels, determined on the basis of the CI user's loudness percepts for pure tones during clinical fitting. These were standardized by conversion to percentage of dynamic range, using the formula

$$y = \frac{x - T}{C - T}, \quad (2)$$

where y is a converted value in percent of dynamic range, x is the stimulation current in units of current levels, T is the threshold level in units of current levels, and C is comfortable level in units of current levels for the stimulated electrode. The capturing procedure was not successful in 6 out of 41 subjects (denoted by asterisks in Table I); captured data therefore encompassed 35 subjects.

MSs were calculated via Fourier transformation of the autocorrelation of the stimulus output patterns (Singh and Theunissen, 2003). Bandpass filtering based on filter-banks using the same cut-off frequencies as in the electrodes in the CI device was first performed using the Nucleus Matlab Toolbox, a part of the NIC© proprietary software (Cochlear, 2002) generously provided by Cochlear AG. MSs were calculated by autocorrelating and then Fourier-transforming each envelope. This analysis included the frequency range between 75 and 225 Hz, which encompassed the F0 values of all stimuli [the upper bound of 225 Hz was one-quarter of the lowest stimulation rate (900 Hz), preventing aliasing effects in the data analysis]. The lower cut-off value of 75 Hz removed the effects of low-frequency modulation due to acoustic variation at syllabic/phonemic levels.

To measure the overall availability of F0 modulation information for a particular stimulus sound in each subject, the MSs of electrode outputs were summed over all channels that (1) were stimulated for at least 25% of the stimulus duration and (2) possessed distinct F0-related peaks in their acoustic MSs (peaks less than 40 Hz [a typical just-noticeable difference (JND) for CI temporal pitch for F0 values around 200 Hz (Zeng, 2002)] from the F0 value). To assess the availability of temporal cues for each stimulus item in each subject, the following algorithm was followed: (1) Find all local peaks in the MS that have energy at 0.8 or more of the maximum amplitude between 75 and 225 Hz and (2) find the closest peak to F0 with relatively greatest intensity. Using this procedure, each stimulus was assigned a CI MS F0-related peak value for each subject.

To assess the influence of the number of electrodes carrying F0-related modulation cues and their relative strength, cue-carrying electrodes were sorted according to the relative strength of their F0-related peaks ("modulation strength"). The analysis used the mean modulation strengths of the four strongest electrodes, and was performed both across response types (correct vs incorrect trials) and gender (male vs female speech items).

Individual DTEs fell into two broad groupings ($<$56 and $>$89 Hz), meaning that subjects with larger and smaller thresholds experienced a different number of trials and a different range of frequency differences. To assess the performance of their CI devices in an unbiased fashion, two different stimulus sets were selected based on the performance of subjects who scored above chance in the identification procedure. Stimulus set 1 (SS1) consisted of 15 stimulus pairs chosen from trials that were correctly responded to by 2 or more performing subjects and that were never incorrectly responded to by any of these subjects. Stimulus set 2 (SS2) contained 15 stimulus pairs chosen from trials responded to incorrectly by at least 2 performing subjects and that were never correctly responded to. The F0 separation between speech items in these trials was similar for the two sets and was limited to the range between 10 and 60 Hz ($\mathrm{mean}_{SS1} = 37.3 \pm 3.5$ Hz, $\mathrm{mean}_{SS2} = 31.7 \pm 3.6$, Mann–Whitney U-test, $Z = -0.975$, $n_{SS1} = 15$, $n_{SS2} = 15$, $p = 0.33$).

Finally, the quality of F0 cues to voice gender was quantified using the MS peak closest to F0, and the centroid position in the MS for each stimulus pair. Linear regressions were calculated using the F0 difference between stimuli as the independent variable, and the difference in CI MS peaks or MS-centroid positions as the dependent variable. The r^2 value for each subject was used as an estimator of the quality of information that each CI processor provided about each type of cue.

III. RESULTS

A. Voice gender identification

1. Behavioral data

Figure 2 shows mean voice gender identification results for the CI participants. Scores between 13 and 26 correct responses are consistent with chance performance based on the binomial distribution. Overall, 18 of the CI participants (44% of the CI subjects) could correctly identify the sex of the speaker at better-than-chance performance levels (mean value of $84.3 \pm 1.0\%$ correct) from an isolated stimulus;

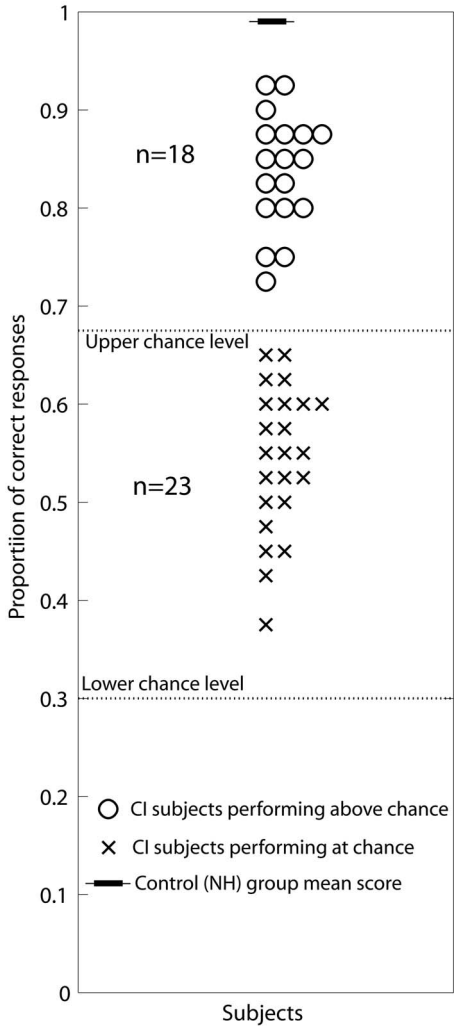

FIG. 2. Proportion of correct responses of 41 subjects with CIs in the fixed identification procedure. Dashed lines indicate the upper and lower boundaries for chance performance set by the binomial distribution. Performing CI subjects achieved a mean of $84.3 \pm 1.0\%$ correct responses; non-performing CI subjects had a mean of $54.4 \pm 2.0\%$ correct responses. The control group of NH subjects performed with a mean accuracy of $98.0 \pm 0.4\%$.

these subjects will subsequently be referred to as "performing" (P) subjects. The other 23 (56% of the study CI population), with a mean value of $54.4 \pm 2.0\%$ correct, were unable to identify the gender of isolated stimulus voices, and will subsequently be referred to as "non-performing" (NP) subjects.[2] All subjects in the control group performed with maximal or near-maximal performance (mean $98.0 \pm 0.4\%$ correct), confirming that the stimulus set was not of abnormal difficulty. Further analysis did not reveal any differences in CI subjects' performance between male and female speech items, independently of whether they could or could not correctly label them.

An analysis by items in subjects with above-chance performance is shown in Fig. 3. This reveals a V-shaped notch in the proportion of correct responses in relation to the speaker's F0, reflecting response uncertainty in the region of gender-ambiguous F0 values.[3]

2. Temporal F0-related modulation cues

The correlations between F0-related peaks measured from CI MSs and the corresponding F0 values measured from the stimulus waveforms were nominally significant in all 35 subjects that CI output signals were successfully captured from (15 performing and 20 non-performing), confirming that all CI devices provided robust F0-related temporal cues. The devices of performing and non-performing subjects did not significantly differ in the magnitudes of these correlations ($\text{mean}_P = 0.84$[SD=0.10], $\text{mean}_{NP} = 0.79$[SD=0.14], Mann–Whitney U-test, $Z = 0.98$, $n_P = 15$, $n_{NP} = 20$, $p = 0.325$). However, over all subjects, the correlation coefficients of correctly identified items were significantly higher than those of incorrectly identified items (means: 0.83 [SD=0.15] vs 0.54 [SD=0.48], Mann–Whitney U-test, $Z = 3.15$, $n = 35$, $p = 0.002$), indicating that the quality of temporal cues provided by the devices was significantly related to perceptual performance.

The 15 performing subjects had a large difference in correlation values between correct and incorrect items (means of 0.87 [SD=0.1] vs 0.21 [SD=0.56], Mann–Whitney U-test, $Z = 3.73$, $n = 15$, $p = 0.0002$). In contrast, there was no difference in the magnitude of the correlation coefficients between correct and incorrect items in the 20 non-performing subjects (means of 0.80 [SD=0.17] vs 0.78 [SD=0.18], Mann–Whitney U-test, $Z = 0.72$, $n = 20$, $p = 0.47$).

To combine the data from both types of voices into a single measure, an analysis of the distance of the F0-related peak in the CI MSs from the F0 "gender boundary" (the half-way point between the mean position of the male and female F0-related peaks of all speech items) and behavioral performance was carried out. All performing subjects had a positive correlation between these two measures, which was significantly different from 0 (Pearson correlation coefficients ranged from 0.33 and 0.83, $p < 0.05$ in all subjects; p-values were calculated using Fisher's r-to-z transform), suggesting that performance was related to how clear male or female temporal F0 cues were.

The acoustic and CI F0 distances of each item from the gender boundary were averaged across subjects, and pooled within performing and non-performing groups (Fig. 4). The F0 modulation information provided by the CI of the two groups is very similar, exhibiting a V-shaped relationship between acoustic and CI F0 distance from the gender boundary (this V-shape was previously observed in the analysis by items of the behavioral data from the performing subjects shown in Fig. 3). A regression of the form $y = B1 + B2 \times x$ between CI F0 distance and the proportion of subjects who correctly responded to each item was calculated for each subject group (where y is the proportion of subjects who responded correctly and x is the frequency distance of the F0-related peak in the CI MS from the gender boundary). The regression was significant and accounted for 24% of the observed gender identification variation in the performing CI subjects [$r^2 = 0.24$, $F(13) = 11.691$, $p = 0.002$], suggesting that they may use additional cues. The regression was not significant for the non-performing subjects [$r^2 = 0.05$, $F(18) = 1.913$, $p = 0.17$].

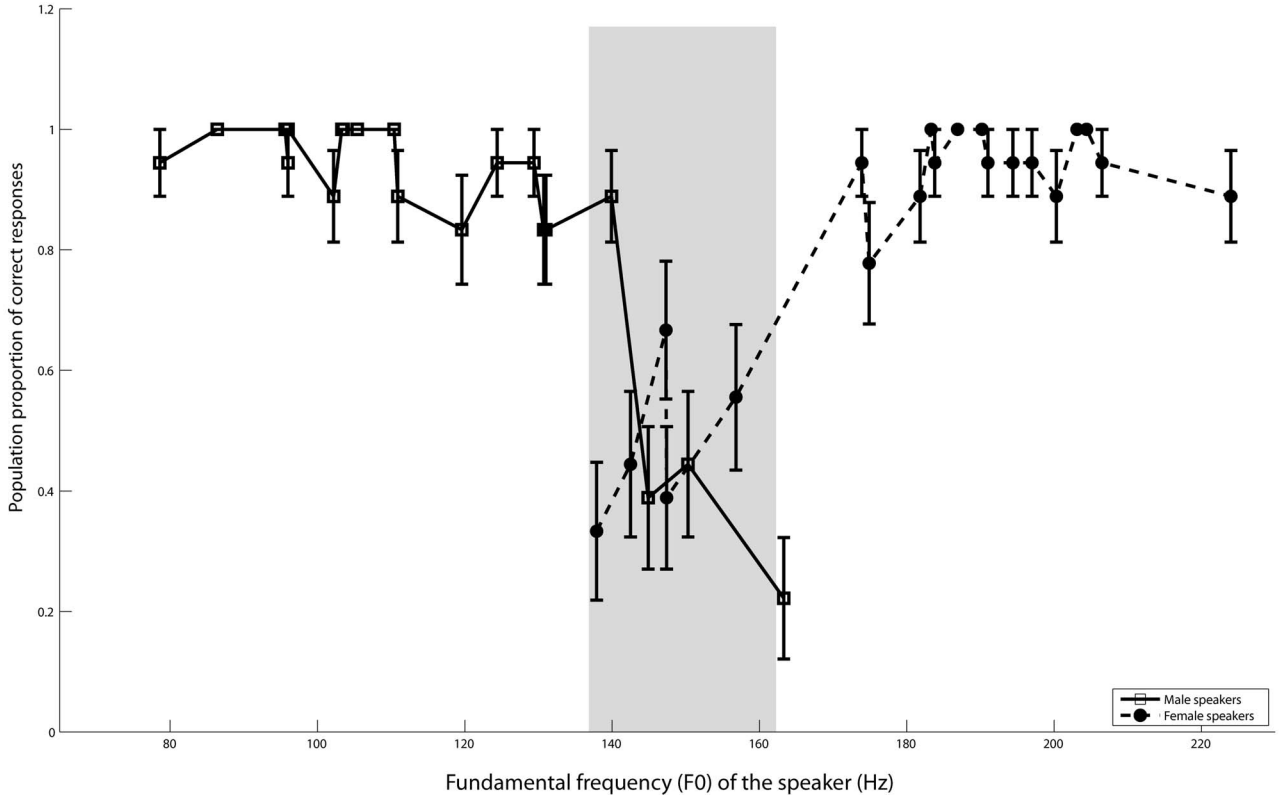

FIG. 3. Proportion of correct responses in performing subjects with CIs sorted by stimulus F0. The shaded area shows the region of overlapping F0 for male and female speech items. Error bars equal one standard error.

3. Number of channels carrying temporal modulation cues

In the gender identification task, the number of active electrodes was different between correctly-responded-to and incorrectly-responded-to trials in only 3 (two performing) of the 35 subjects from whom the capturing procedure was successful. Similar results were obtained with trials sorted according to the gender of the speaker (two subjects showed a significant difference in electrode numbers between male and female speech items). While a few subjects may show overall differences in the number of electrodes activated by male and female voices, it appears that this parameter did not play a general role in the identification task.

4. Magnitude of temporal modulation cues

The mean positions of electrodes with the four largest MS strength values were significantly different between female and male speech items in 4 (all performing) out of the 35 subjects. This suggests that in these four subjects there might be a relationship between the average position of the four electrodes having maximal MS strengths and voice F0. Two of these subjects (with relatively high performance) had a significant linear regression of MS strength on F0 with a negative slope for correctly-responded-to trials [subject CI24: $y=21.298-0.0042282X$, r^2(adjusted)$=0.18$, $F(30)=7.99$, $p=0.008$; subject CI31: $y=21.406-0.0050484X$, r^2(adjusted)$=0.17$, $F(33)=7.73$, $p=0.009$). Such an F0-MS strength relationship was not present for the other two subjects, who showed a greater variability of the mean electrode positions (ranges: 7.75–20 electrode units) compared to the former group (ranges: 19.5–21 electrode units). Still, on average, all four subjects had a systematic shift in the mean positions of the four electrodes containing the largest modulations between male and female items.

34 of the 35 subjects whose CI data were recorded showed a significant difference in the mean amount of modulation strength of the four most "active" electrodes between female and male speech items. Further analysis revealed a clear dependence of the strength of F0-related modulation and the value of F0. These equations were fitted with an exponential regression of the form $\mathrm{MS}_{\mathrm{strength}}=a\times e^{b\times F0}$. All 35 subjects showed a significant regression and there were no differences in the magnitude of the coefficients of determination between performing and non-performing subjects ($\mathrm{mean}_{\mathrm{P}}=0.54$[SD$=0.11$], $\mathrm{mean}_{\mathrm{NP}}=0.53$[SD$=0.13$], Mann–Whitney U-test, $Z=0.15$, $n_{\mathrm{P}}=15$, $n_{\mathrm{NP}}=20$, $p=0.88$). When this analysis was performed between correctly-responded-to and incorrectly-responded-to items, only one (non-performing) subject showed a difference in modulation strengths (means: 3.53×10^{-6} vs 6.11×10^{-6}, Mann–Whitney U-test, $Z=1.98$, n(correct)$=24$, n(incorrect)$=15$, $p<0.05$). In general, it seems that the subjects did not utilize these modulation strength differences while performing gender identification, even though speech items with lower F0s produced stronger F0-related modulations in all CI devices.

5. Place cues

The mean male-female spatial distance for all participating subjects was 0.28 ± 0.03 electrode units and is significantly different from 0 (Wilcoxon matched-pair sign rank

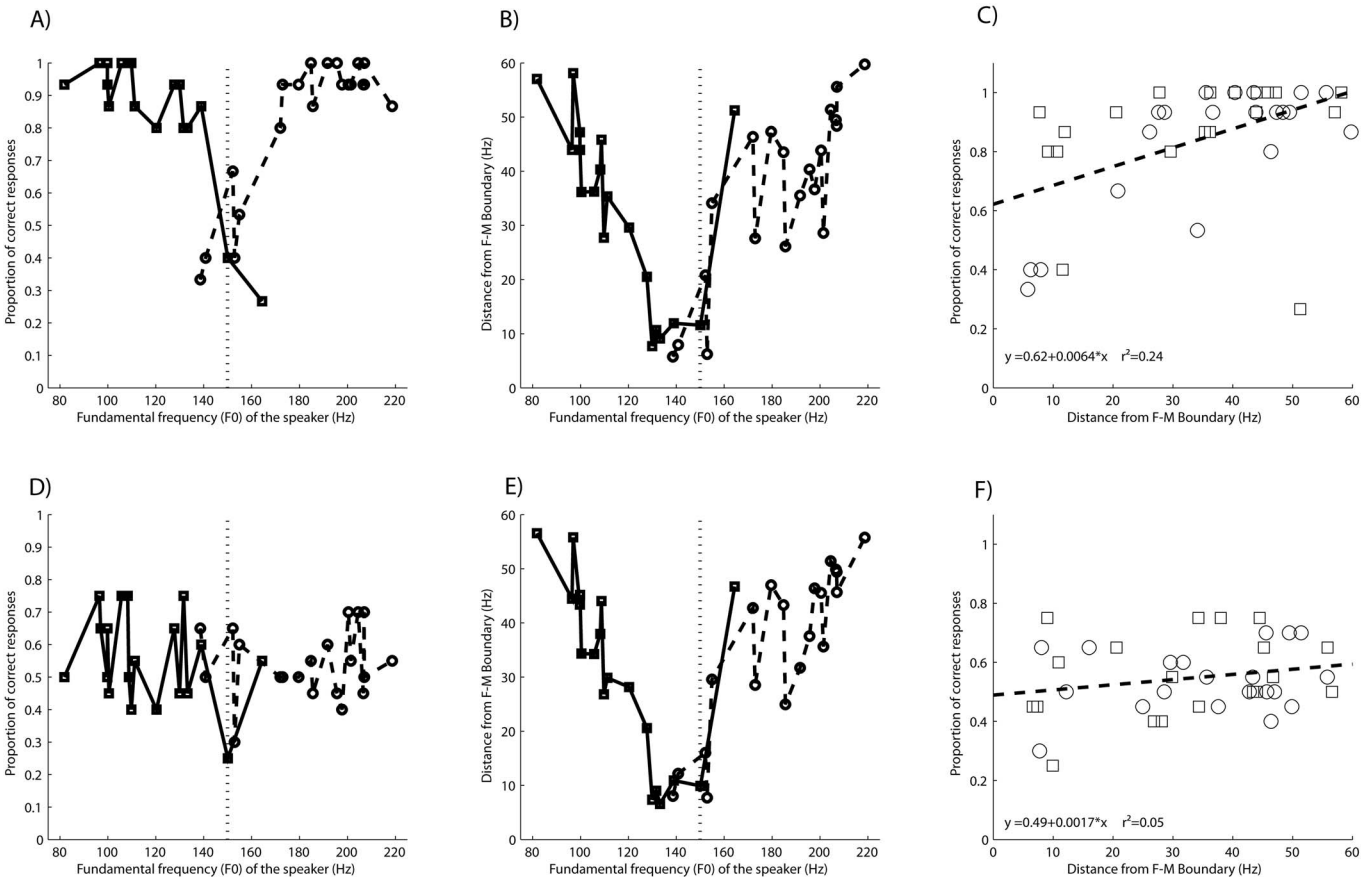

FIG. 4. The availability of F0 temporal cues provided by the CI as a function of F0, and its relation to behavioral performance. Upper panels (A)–(C) show data for the performing group of subjects, while lower panels (D)–(F) show results for the non-performing subjects. In all panels, circles denote female speech items, and squares denote male speech items. Panels (A) and (D) show the relationship between the proportions of subjects who responded correctly and the F0 of the speech item. Panels (B) and (E) show average F0 distances from the gender boundary as recorded from the CI output patterns for each speech item. Panels (C) and (F) show the relationship between the proportion of subjects who correctly identified each speech item and the corresponding F0 distance from the gender boundary as recorded from the CI output patterns. The dashed line represents the linear regression. Coefficients of linear regression and their corresponding coefficients of determination are given on the bottom left of the panel.

test, $T^+=594$, $Z=-4.57$, $n=35$, $p<0.001$). The mean spatial difference in the devices of performing group subjects was statistically the same as the difference in the devices of non-performing subjects (Mann–Whitney U-test, $Z=-0.1167$, $n_P=15$, $n_{NP}=20$, $p=0.91$). Although place cues were available to all subjects, they were of small magnitude. An additional analysis using only the ten most extreme F0 values in each gender group still showed no difference in the magnitude of place cues between the devices of performing and non-performing subjects (Mann–Whitney U-test, $Z=-0.38$, $n_P=15$, $n_{NP}=20$, $p=0.70$). Analyzing place data by response type (correctly-responded-to items and incorrectly-responded-to items) again revealed no differences in the devices of either group of CI subjects (performing group, all items: U-test, $Z=0.55$, $n_C=15$, $p_C=0.59$; non-performing group, all items: U-test, $Z=0.47$, $n_I=20$, $p_I=0.64$). Place cue distances from the gender boundary were not associated with acoustic F0 frequency distances in any of the subjects. While subtle information about center-of-mass place cues is available in the devices of all subjects, it seems to be of little utility for voice gender identification.

Place cues may provide CI subjects with additional information in the form of consistent covariation with temporal cues, which could be masked by the stronger salience of temporal information. However, a joint analysis assessing the relationship between temporal (F0-related peak in MS) and spectral (CM positions of the stimulus output pattern) distances from the gender boundary for all subjects revealed no significant relationships between place and temporal cues (analysis not shown).

B. Adaptive speech-based F0 discrimination of voice gender

1. Behavioral data

The control group performed perfectly up to the limits of the experimental stimulus pairs (12 subjects did not show reversals, with the remaining 3 subjects showing two reversals only at $\Delta F0<0$). This demonstrates that the task works well for hearing children. The picture is markedly different for the CI users. Figure 5 shows the DTEs for each subject. Two groups of subjects are readily apparent: (i) 23 CI children produced small ("good") DTEs with values of 55.8 Hz or less, and (ii) the other 18 CI subjects had substantially higher ("worse") DTEs (89.8 Hz and higher). Threshold in this testing scheme corresponds to a performance of 70.7% correct.

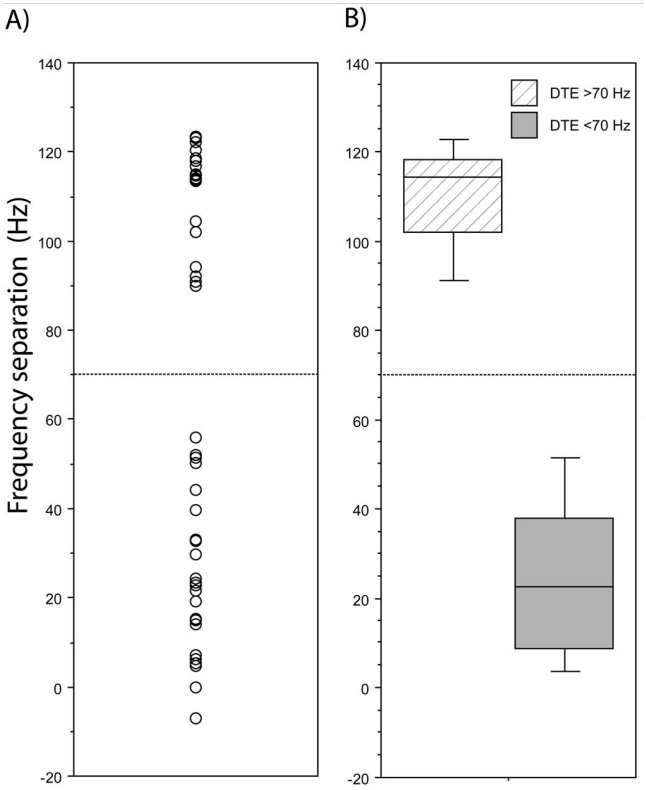

FIG. 5. Scores in the adaptive speech-based F0 discrimination procedure: (A) adaptive DTEs for each CI subject; (B) box plot of the DTE for both performance groups. In-box horizontal lines are the median values. Plotting bar boundaries are between the 25%th and 75%th percentiles; error bars go between the 10%th and 90%th percentiles.

2. Temporal cues

Performing and non-performing subjects[4] had similar r^2 values for both measures of F0 information in both stimulus sets [Mann–Whitney U tests, stimulus set 1: r^2(MS-centroid)$=0.59\pm0.03$ vs 0.54 ± 0.04, $Z=-1.20$, $n_P=19$, $n_{NP}=14$, $p=0.23$; r^2(F0-peak)$=0.58\pm0.04$ vs 0.54 ± 0.07, $Z=-0.22$, $n_P=19$, $n_{NP}=14$, $p=0.82$; stimulus set 2: r^2(MS-centroid)$=0.31\pm0.04$ vs 0.35 ± 0.04, $Z=-0.49$, $n_P=19$, $n_{NP}=14$, $p=0.62$; r^2(F0-peak)$=0.22\pm0.03$ vs 0.21 ± 0.05, $Z=-0.26$, $n_P=19$, $n_{NP}=14$, $p=0.8$]. This empirically confirms that the CI devices of performing and non-performing subjects provided similar F0 information.

Figure 6 shows differences between the stimulus sets in the quality of the information transmitted by the CI. Items from stimulus set 1 provide significantly better temporal information than items from stimulus set 2 in all subjects (Wilcoxon matched-pair sign rank test for F0-peak and MS-centroid information: $T^+=538$, $Z=-4.60$, $n=33$, $p<0.0001$). These differences exist in both performance groups for both temporal measures (Wilcoxon matched-pair sign rank test, F0-peak information; performing group, $T^+=189$, $Z=-3.78$, $n=19$, $p=0.0002$; non-performing: $T^+=96$, $Z=-2.73$, $n=14$, $p=0.0063$; MS-centroid information: performing group, $T^+=186$, $Z=-3.66$, $n=19$, $p=0.0003$; non-performing: $T^+=96$, $Z=-2.73$, $n=14$, $p=0.0063$).

Another way to look at the quality of temporal information provided by the CI is the degree of consistency between the CI MS spectra peaks and MS-centroid positions, as measured by their correlation. Over all subjects, the correlation coefficient was nominally larger in stimulus set 1 compared to stimulus set 2 [0.57 ($n=33$, $p<0.001$, Fisher r-to-z) vs 0.39 ($n=33$, $p=0.02$, Fisher r-to-z)], but the difference in the magnitude of these two correlation coefficients was not significant ($z=0.91$, $p=0.35$). Similar results were found when this analysis was repeated separately for each performance group (analysis not shown).

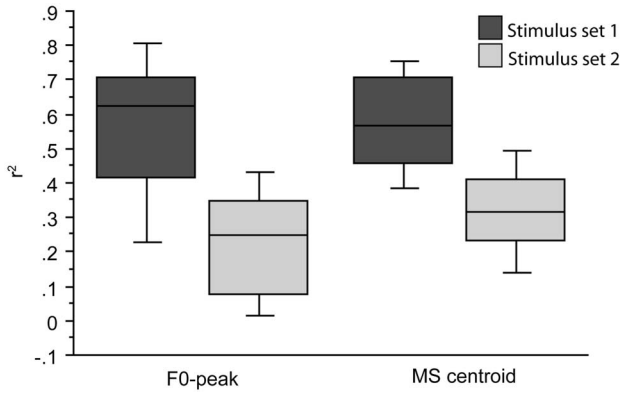

FIG. 6. Differences in r^2 between stimulus sets for measures of temporal information provided by the CI device (F0-related peak, left; and MS-centroid, right). In-box horizontal lines are the median values. Plotting bar boundaries are between the 25%th and 75%th percentiles; error bars go between the 10%th and 90%th percentiles.

In order to assess whether either of the temporal measures provided better cues than the other, a comparison of the r^2 values for items from stimulus set 1 was examined, but this revealed no advantage over all subjects, or for the performing or non-performing groups. The same analysis for stimulus set 2 revealed that MS-centroids were represented with a higher degree of fidelity compared to F0-peak information [Wilcoxon signed rank test, performing group: r^2(MS-centroid)$=0.31\pm0.04$ vs r^2(F0-peak)$=0.22\pm0.03$, $T^+=41$, $Z=-2.173$, $n=19$, $p=0.03$; non-performing group: r^2(MS-centroid)$=0.35\pm0.04$ vs r^2(F0-peak)$=0.21\pm0.05$, $T^+=18$, $Z=-2.166$, $n=14$, $p=0.03$; all subjects: r^2(MS-centroid)$=0.33\pm0.03$ vs r^2(F0-peak)$=0.22\pm0.03$, $T^+=112$, $Z=-3.011$, $n=33$, $p=0.003$). The fact that MS-centroid information is better-represented in stimulus items that CI subjects discriminate more poorly suggests that CI users may not be able to extract information from this cue very effectively.

Adaptive discrimination performance could be limited by the availability of relevant temporal information. However, an examination of regressions between the r^2 values of MS-centroid and F0-peak temporal information and adaptive DTEs yielded no significant relationships.

3. Place cues

The analysis of place cues in the form of the stimulus pair differences between center-of-mass position of the stimulus output patterns and the corresponding F0 revealed that only two subjects had significant correlations for correctly-responded items (performing subject CI30, correlation coefficient=0.40, $p=0.02$, and $n=34$; non-performing subject CI25, correlation coefficient=0.48, $p=0.05$, and n

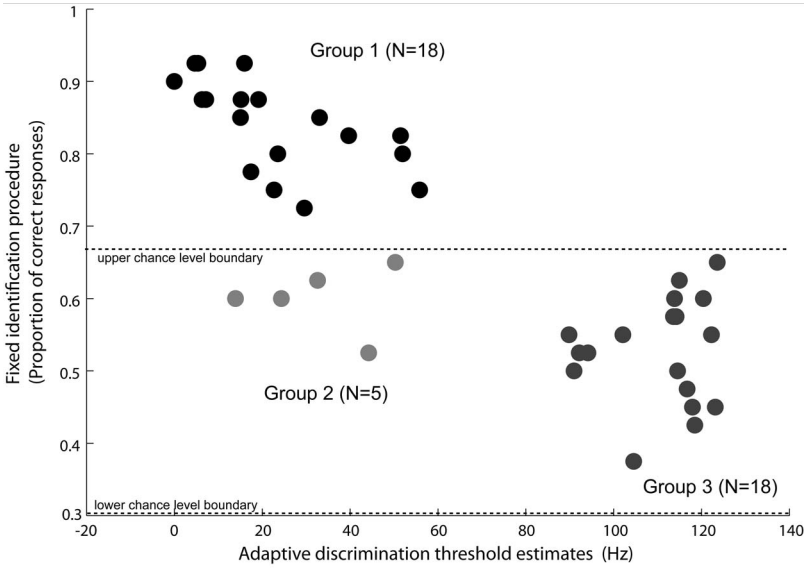

FIG. 7. Two-dimensional plot of the joint results of the fixed identification and adaptive speech-based F0 discrimination procedures. Each data point represents one subject.

=17). As in the identification task, temporal cues in the form of amplitude modulation of pulse trains appear to be a major source of useful information about speaker gender that CI devices are transmitting.

C. Joint results of the fixed identification and adaptive discrimination procedures

Figure 7 illustrates the joint results of the two procedures used in this study. Over all subjects, there is strong correlation between the performances obtained from the two different assessments (correlation=−0.822, $n=41$, $p<0.0001$, Fisher r-to-z). The data indicate three natural groupings of CI subjects based on their task performance. Group 1 (upper left) consists of 18 subjects with above-chance performance in voice gender identification and adaptive DTEs below 70 Hz. These subjects were able to identify gender at above-chance levels and performed at 70.7% or better in the adaptive discrimination task. Group 2 consists of five subjects (lower left) who were unable to correctly identify voice gender in spite of having speech-based F0 DTEs between 15 and 60 Hz, overlapping completely with many of the subjects in group 1. Group 3 consists of 18 subjects who were not able to correctly identify voice gender, and showed poor adaptive speech-based F0 DTEs as well.

One possibility for the differential performance in the two experiments could be that the subjects from group 2 learned to identify voice gender in the course of the identification procedure (which always preceded the discrimination experiment, and had feedback to maintain the subject's interest). However, none of these subjects showed significant differences in performance for the initial and final blocks of speech items in the identification procedure when assessed with Mann–Whitney tests, suggesting that learning did not occur.

IV. DISCUSSION

This study evaluated how children with CIs performed in gender identification tasks using naturalistic stimuli, and empirically assessed the quality of information that their implant devices were providing to them. In the fixed identification procedure, 18 out of 41 subjects (44%) could identify gender from 2-s speech items taken from a set representing 20 male and 20 female speakers, whereas the other 23 (56%) of the subjects were unable to perform this task. The results demonstrate that individual variation in the use of CI information is not only seen in complex cognitive tasks such as speech perception (Blamey *et al.*, 2001), but is also observed in more basic perceptual tasks such as identification of voice gender.

There are several exogenous and endogenous factors that might contribute to this variation. Major exogenous factors include technical and clinical aspects of cochlear implantation (the device and the surgery), whereas the major endogenous factors include the fundamental properties of the auditory system together with the developmental trajectory of deafness and CI use. A major but little-addressed issue examined by the present study concerns the quality of information that CI devices provide to their users.

Patterns of stimulation delivered by each subject's CI device were examined, and no general difference was found in the quality of information provided by the devices of subjects who perform gender identification well and those who do not; performance differences appear to result from the ability of CI users to make use of these cues. This variation in the ability to use information may result from differential effects of electrode position and insertion depth, and interactions with cochlear biophysics (which are likely to be different in different subjects), resulting in differences in the number, location, and quality of stimulation of auditory nerve fibers. All the CI users in the present study underwent full-depth electrode insertion, and it was not possible to compare the relatively small electrode insertion depth differences among them with voice gender identification performance.

However, data were collected on exogenous variables such as age of the CI surgery, age of the onset of deafness, duration of CI use, and duration of deafness; an evaluation of the role of these variables will appear in a separate publication.

For the group of CI wearers who were proficient at gender identification, the quality of temporal information was significantly better for the items that they responded to correctly when compared to incorrect items. Such a contrast was not found in the group of subjects who performed poorly in gender identification, suggesting that the performing group of subjects could use the temporal information provided by their devices. Place cues were available but of small magnitude, with no evidence for their use. A greater use of temporal cues over spectral cues has also been suggested by previous researchers. For example, Fu *et al.* (2004) compared gender discrimination abilities using vowels, and found up to 30% better performance when temporal cues were available. Fu *et al.* (2005) compared gender identification abilities with vowels spoken by talkers coming from two sets, with an F0 separation between female and male speakers of 100 and 10 Hz. CI users performed well with the first talker set (F0 separation of 100 Hz), but their performance decreased sharply for the set of speakers with small F0 separation.

The issue remains whether the non-utilization of place information in the experiments presented here is due to difficulties in detection, or masking by the stronger presence of temporal cues. A preference for temporal over place cues for fundamental frequency was confirmed by Laneau and Wouters (2004b), who found an increase in just-noticeable differences for F0 discrimination in four CI users from just above 1 octave when place cues were presented alone, to 6%–60% of an octave when temporal information was added.

The fact that F0 information only accounts for approximately one-quarter of the variation in voice gender identification [Fig. 4(C)] suggests the use of additional gender-related cues. One such cue was present in the form of variation in electrode modulation strength. All subjects' devices exhibited a significant relationship between the modulation strength of the four most active electrodes and voice F0, but subjects appeared insensitive to it. It is unclear whether modulation strength cues are not useful because they are perceptually difficult to detect, because CI users are unable to interpret them, or because users are not paying attention to them; it is equally unclear whether CI users could learn to make use of them. Further research with CI subjects and NH subjects listening to CI simulations is necessary to better understand and exploit any perceptual relevance of electrode modulation strengths and their spatial patterning for sound processing schemes.

The CI speech coding algorithm implemented in the devices studied here does not appear to provide equally good temporal representations of fundamental frequencies for all voices. Stimulus pairs that were more problematic for gender identification also had poorer temporal representations of fundamental frequency. Since CI output patterns are the result of a complex interplay between clinical parameter settings of the devices and spectro-temporal features of the incoming sounds, this suggests the need for a better optimization strategy for the speech coding algorithm in order to facilitate gender identification for a wider variety of voices.[5]

Finally, this study identified five CI users who could not identify the gender of singly-presented voices but could do so at better-than-chance levels when two voices were sequentially presented. This dissociation of abilities for voice gender identification and adaptive speech-based F0 discrimination is striking, as these users had periods of CI experience (an average of 4.04 ± 0.96 years of use) equivalent to the 18 CI users who could successfully perform both tasks (an average of 4.02 ± 1.06 years of use).

The two paradigms differ in terms of how subjects might strategically perform them. Auditory events appear to produce a transient auditory sensory memory trace lasting up to 10–20 s (Cowan, 1984; Semal and Demany, 1991; Krumhansl and Iverson, 1992; Clement *et al.*, 1999; Caclin *et al.*, 2006). In a single-interval gender identification task, gender-relevant acoustic information in a stimulus item must be compared with long-term memory representations of gender-related acoustic features. Two items presented one after the other within ~5 s could be compared without reference to long-term representations. Winkler *et al.* (2002) and Winkler and Cowan (2005) assessed ways in which short-term auditory ("surface") information gets encoded into longer-term memory, suggesting that the auditory context within which the sounds are to be remembered plays an important role. Perhaps some general types of auditory experiences at an early age are necessary to set up relations between sounds and contexts that facilitate the formation of long-term categorical auditory memories. This issue has important implications for the ease with which CI users may be able to use different kinds of auditory information, and needs to be examined more rigorously in future research.

V. CONCLUSION

Forty-one juvenile cochlear implantees were exposed to naturalistic speech samples from a variety of speakers, and asked to perform two voice gender perception tasks; the stimulus output patterns of their CIs were also documented for each speech sample. CI electrical output features related to voice fundamental frequency (F0) and spectral envelope were evaluated in relation to behavioral performance. Temporal and place cues were equally available in all CI devices, but only about half of the subjects were able to label gender correctly. Subjects who could identify voice gender appeared to utilize temporal cues but showed no evidence of using place cues. At least one other robust F0-related cue was present in the output of all CI devices that participants appeared unable to make use of. A subgroup of participants could discriminate voice gender when two contrasting voices were presented in succession, but were unable to identify gender when voices were singly presented, suggesting that it may be fruitful to more carefully examine the characteristics of long-term auditory category formation in CI user populations.

ACKNOWLEDGMENTS

This work was supported by the Croatian Ministry of Science, Education and Sports Grant (Grant No. 207-0000000-2293), the Central European Initiative Science and Technology Network Research Fellowship to D.K., NSERC Grant No. 298612 and CFI Grant No. 9908 to E.B., and SISSA. The authors thank Ernst von Wallenberg and Cochlear AG, Basel for providing hardware, Andrew Vandali (CRC HEAR, Melbourne) and Peter Seligman (Cochlear Ltd., Melbourne) for software and consultation, Davor Petrinović (Faculty of Electrical Engineering and Computing, Zagreb) and Vladimir Kozina (AKO Electrical Engineering, Zagreb) for technical support and the staff of SUVAG Polyclinic for logistic support, and Colette McKay and two anonymous reviewers for helpful comments.

APPENDIX: COCHLEAR IMPLANT MICROPHONE ISOLATION

For subjects using Sprint processors, an Intra-Operative direct cable (manufactured by Cochlear Ltd., Sydney, Australia and provided by Cochlear AG, Basel, Switzerland) that shuts off the external microphone was used during these experiments. However, most of the subjects used an Esprit 3G CI processor (typically programed with the mixing mode enabled). Direct analog input to the processor via a cable connection would not normally switch off the external processor's microphone, resulting in the analog input being mixed with environmental sounds. This was undesirable during the sound measurements and experimental stimulus presentations, and it was equally undesirable to change any parameter settings on a subject's processor. Therefore, an acoustic isolation chamber for the implant ear-piece was constructed. This consisted of a small thermos double-walled container filled with cement aggregated with small iron balls (1–2 mm in diameter) in order to obtain high density for reducing sound transmission. The Esprit 3G processor was placed within the insulated box, which was then sealed. A 30-cm long coil cable was connected to the processor. The insulated box was suspended in the vicinity of and behind the subject's head, making sure that it did not distract the subject or interfere with their comfort.

The attenuation of the insulated box was measured with an internal microphone (ECM 335-62M frequency range 20–20 000 Hz) and an external source of white noise [95 dB SPL (*C*-weighting, slow reading)] using three different frequency ranges: (1) octave-wide filtered white noise (60 dB/octave roll-off) centered at 1000 Hz (27 dB attenuation), (2) octave-wide filtered white noise (60 dB/octave roll-off) centered at 2000 Hz (25 dB attenuation), and (3) stimulation by a loudspeaker with a frequency range 100–20 000 Hz (30 dB attenuation). In addition, a generic CI device (Esprit 3G, 15 active electrodes, frequency table 7, frequency range 120–8658 Hz, and stimulation rate: 900 Hz pps/channel) sealed inside the box required augmentation of the sound by 36 dB to produce similar average pulse output activity.

[1]The F0 values of the speech items used here range from 80 to 220 Hz; CI maps typically have this entire frequency range assigned to the most apical active electrode. How is it possible to use any F0 place cue information when it all falls within one channel? F0 differences are also correlated with differences in the distribution of spectral envelopes between male and female speakers (Ives *et al.*, 2005; Smith and Patterson, 2005; Smith *et al.*, 2005, 2007). In the course of auditory experience with the device, CI users might build up a representation of the population distribution of the center-of-mass of stimulation over the electrode array for male and female voices, which could allow them to discern stimulation patterns whose center-of-mass differs by fractions of an electrode position.

[2]If the male and female voices with an overlapping F0 are excluded from the analysis, the mean performance increased significantly to $94.1 \pm 2.0\%$ in the performing group of CI subjects (Mann–Whitney U-test, $n=18$, $Z=-4.02$, and $p<0.000\,01$), but was unchanged in the non-performing group of CI subjects (mean performance $56.0 \pm 2.0\%$, Mann–Whitney U-test, $n=23$, $Z=-0.26$, and $p=0.79$).

[3]One male speech item was consistently ambiguous in the control group, and this item was excluded from all further analyses.

[4]CI output patterns captured from stimulus sets 1 and 2 were unusable in two non-performing subjects because of equipment failure.

[5]Cleary and Pisoni (2002) found that CI users performed poorly in voice identification when the linguistic content of paired utterances differed. This does not necessarily conflict with the present findings. In Cleary and Pisoni, 2002, the subjects performed a "same-different" task for identifying the speaker (using 3 different female voices) in the face of variation in the content of utterances, in contrast to the voice gender identification task using 20 male and 20 female voices studied here. Short-term comparison may have facilitated the performance of the subjects in the present study, but not those in Cleary and Pisoni, 2002, because gender identification can be successfully based on a few relatively static vocal cues (even with a large number of voices). In contrast, useful cues for speaker identification include more dynamic voice features, and subjects need to use different combinations of static and dynamic features to tell the difference between different types of voices. Short-term comparisons would help with the former kind of task, but may hinder overall performance in the latter kind of task.

Amitay, S., Irwin, A., Hawkey, D. J., Cowan, J. A., and Moore, D. R. (**2006**). "A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners," J. Acoust. Soc. Am. **119**, 1616–1625.

Bachorowski, J. A., and Owren, M. J. (**1999**). "Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech," J. Acoust. Soc. Am. **106**, 1054–1063.

Baker, R. J., and Rosen, S. (**2001**). "Evaluation of maximum-likelihood threshold estimation with tone-in-noise masking," Br. J. Audiol. **35**, 43–52.

Blamey, P. J., Sarant, J. Z., Paatsch, L. E., Barry, J. G., Bow, C. P., Wales, R. J., Wright, M., Psarros, C., Rattigan, K., and Tooher, R. (**2001**). "Relationships among speech perception, production, language, hearing loss, and age in children with impaired hearing," J. Speech Lang. Hear. Res. **44**, 264–285.

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

Boersma, P., and Weenink, D. (**2006**). *Praat: Doing Phonetics by Computer (Version 4.4.32)* [Computer program]. Retrieved October 1, 2006 from http://www.praat.org/.

Burns, E. M., and Viemeister, N. F. (**1976**). "Non-spectral pitch," J. Acoust. Soc. Am. **60**, 863–869.

Burns, E. M., and Viemeister, N. F. (**1981**). "Played-again SAM: Further observations on the pitch of amplitude-modulated noise," J. Acoust. Soc. Am. **70**, 1655–1660.

Caclin, A., Brattico, E., Tervaniemi, M., Naatanen, R., Morlet, D., Giard, M. H., and McAdams, S. (**2006**). "Separate neural processing of timbre dimensions in auditory sensory memory," J. Cogn. Neurosci. **18**, 1959–1972.

Carlyon, R. P., and Deeks, J. M. (**2002**). "Limitations on rate discrimination," J. Acoust. Soc. Am. **112**, 1009–1025.

Chang, Y. P., and Fu, Q. J. (**2006**). "Effects of talker variability on vowel recognition in cochlear implants," J. Speech Lang. Hear. Res. **49**, 1331–1341.

Childers, D. G., and Wu, K. (**1991**). "Gender recognition from speech. Part II: Fine analysis," J. Acoust. Soc. Am. **90**, 1841–1856.

Cleary, M., and Pisoni, D. B. (**2002**). "Talker discrimination by prelingually

deaf children with cochlear implants: Preliminary results," Ann. Otol. Rhinol. Laryngol. Suppl. **189**, 113–118.
Cleary, M., Pisoni, D. B., and Kirk, K. I. (**2005**). "Influence of voice similarity on talker discrimination in children with normal hearing and children with cochlear implants," J. Speech Lang. Hear. Res. **48**, 204–223.
Clement, S., Demany, L., and Semal, C. (**1999**). "Memory for pitch versus memory for loudness," J. Acoust. Soc. Am. **106**, 2805–2811.
Cochlear (**2002**). "Nucleus implant communicator (NIC) system overview," Cochlear Ltd.
Cohen, L. T., Busby, P. A., Whitford, L. A., and Clark, G. M. (**1996**). "Cochlear implant place psychophysics 1. Pitch estimation with deeply inserted electrodes," Audiol. Neuro-Otol. **1**, 265–277.
Cowan, N. (**1984**). "On short and long auditory stores," Psychol. Bull. **96**, 341–370.
Drennan, W. R., and Pfingst, B. E. (**2006**). "Current-level discrimination in the context of interleaved, multichannel stimulation in cochlear implants: Effects of number of stimulated electrodes, pulse rate, and electrode separation," J. Assoc. Res. Otolaryngol. **7**, 308–316.
Fu, Q. J., Chinchilla, S., and Galvin, J. J. (**2004**). "The role of spectral and temporal cues in voice gender discrimination by normal-hearing listeners and cochlear implant users," J. Assoc. Res. Otolaryngol. **5**, 253–260.
Fu, Q. J., Chinchilla, S., Nogaki, G., and Galvin, J. J. (**2005**). "Voice gender identification by cochlear implant users: The role of spectral and temporal resolution," J. Acoust. Soc. Am. **118**, 1711–1718.
Gonzalez, J., and Oliver, J. C. (**2005**). "Gender and speaker identification as a function of the number of channels in spectrally reduced speech," J. Acoust. Soc. Am. **118**, 461–470.
Holt, R. F., and Kirk, K. I. (**2005**). "Speech and language development in cognitively delayed children with cochlear implants," Ear Hear. **26**, 132–148.
Ives, D. T., Smith, D. R. R., and Patterson, R. D. (**2005**). "Discrimination of speaker size from syllable phrases," J. Acoust. Soc. Am. **118**, 3816–3822.
Krumhansl, C. L., and Iverson, P. (**1992**). "Perceptual interactions between musical pitch and timbre," J. Exp. Psychol. Hum. Percept. Perform. **18**, 739–751.
Laneau, J., and Wouters, J. (**2004a**. "Multichannel place pitch sensitivity in cochlear implant recipients," J. Assoc. Res. Otolaryngol. **5**, 285–294.
Laneau, J., and Wouters, J. (**2004b**). "Relative contributions of temporal and place pitch cues to fundamental frequency discrimination in cochlear implantees," J. Acoust. Soc. Am. **116**, 3606–3619.
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.
McDermott, H. J., and McKay, C. M. (**1994**). "Pitch ranking with nonsimultaneous dual-electrode electrical stimulation of the cochlea," J. Acoust. Soc. Am. **96**, 155–162.
McKay, C. M. (**2005**). "Spectral processing in cochlear implants," in *Auditory Spectral Processing*, edited by M. S. Malmierca and D. R. F. Irvine (Elsevier Academic, San Diego, CA), pp. 473–509.
McKay, C. M., and McDermott, H. J. (**2000**). "Place and temporal cues in pitch perception: Are they truly independent?," ARLO **1**, 25–30.
McKay, C. M., McDermott, H. J., and Clark, G. M. (**1994**). "Pitch percepts associated with amplitude-modulated current pulse trains in cochlear implantees," J. Acoust. Soc. Am. **96**, 2664–2673.
McKay, C. M., McDermott, H. J., and Clark, G. M. (**1995**). "Pitch matching of amplitude-modulated current pulse trains by cochlear implantees: The effect of modulation depth," J. Acoust. Soc. Am. **97**, 1777–1785.
Moulines, E., and Laroche, J. (**1995**). "Non-parametric techniques for pitch-scale and time-scale modification of speech," Speech Commun. **16**, 175–205.
Nicholas, J. G., and Geers, A. E. (**2007**). "Will they catch up? The role of age at cochlear implantation in the spoken language development of children with severe to profound hearing loss," J. Speech Lang. Hear. Res. **50**, 1048–1062.
Owren, M. J., Berkowitz, M., and Bachorowski, J. A. (**2007**). "Listeners judge talker sex more efficiently from male than from female vowels," Percept. Psychophys. **69**, 930–941.
Semal, C., and Demany, L. (**1991**). "Dissociation of pitch from timbre in auditory short-term memory," J. Acoust. Soc. Am. **89**, 2404–2410.
Singh, N. C., and Theunissen, F. E. (**2003**). "Modulation spectra of natural sounds and ethological theories of auditory processing," J. Acoust. Soc. Am. **114**, 3394–3411.
Smith, D. R. R., and Patterson, R. D. (**2005**). "The interaction of glottal-pulse rate and vocal-tract length in judgments of speaker size, sex, and age," J. Acoust. Soc. Am. **118**, 3177–3186.
Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., and Irino, T. (**2005**). "The processing and perception of size information in speech sounds," J. Acoust. Soc. Am. **117**, 305–318.
Smith, D. R. R., Walters, T. C., and Patterson, R. D. (**2007**). "Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled," J. Acoust. Soc. Am. **122**, 3628–3639.
Spahr, A. J., and Dorman, M. F. (**2004**). "Performance of subjects fit with the Advanced Bionics CII and Nucleus 3G cochlear implant devices," Arch. Otolaryngol. Head Neck Surg. **130**, 624–628.
Stacey, P. C., Fortnum, H. A., Barton, G. R., and Summerfield, A. Q. (**2006**). "Hearing-impaired children in the United Kingdom, I: Auditory performance, communication skills, educational achievements, quality of life, and cochlear implantation," Ear Hear. **27**, 161–186.
Svirsky, M. A., Robbins, A. M., Kirk, K. I., Pisoni, D. B., and Miyamoto, R. T. (**2000**). "Language development in profoundly deaf children with cochlear implants," Psychol. Sci. **11**, 153–158.
Svirsky, M. A., Teoh, S. W., and Neuburger, H. (**2004**). "Development of language and speech perception in congenitally, profoundly deaf children as a function of age at cochlear implantation," Audiol. Neuro-Otol. **9**, 224–233.
Waltzman, S. B., Scalchunes, V., and Cohen, N. L. (**2000**). "Performance of multiply handicapped children using cochlear implants," Am. J. Otol. **21**, 329–335.
Winkler, I., and Cowan, N. (**2005**). "From sensory to long-term memory—Evidence from auditory memory reactivation studies," J. Exp. Psychol. **52**, 3–20.
Winkler, I., Korzyukov, O., Gumenyuk, V., Cowan, N., Linkenkaer-Hansen, K., Ilmoniemi, R. J., Alho, K., and Naatanen, R. (**2002**). "Temporary and longer term retention of acoustic information," Psychophysiology **39**, 530–534.
Wu, K., and Childers, D. G. (**1991**). "Gender recognition from speech. Part I: Coarse analysis," J. Acoust. Soc. Am. **90**, 1828–1840.
Zeng, F. G. (**2002**). "Temporal pitch in electric hearing," Hear. Res. **174**, 101–106.