Title

Speaker recognition with temporal cues in acoustic and electric hearing

Permalink
https://escholarship.org/uc/item/4p71m04k

Journal
Journal of the Acoustical Society of America, 118(2)

ISSN
0001-4966

Authors
Vongphoe, M
Zeng, F G

Publication Date
2005-08-01

Peer reviewed

Speaker recognition with temporal cues in acoustic and electric hearing[a)]

Michael Vongphoe[b)] and Fan-Gang Zeng[c)]
Hearing and Speech Research Laboratory, Departments of Anatomy and Neurobiology, Biomedical Engineering, Cognitive Sciences, and Otolaryngology—Head and Neck Surgery, University of California, Irvine, California 92697-1275

(Received 6 January 2004; revised 6 May 2005; accepted 6 May 2005)

Natural spoken language processing includes not only speech recognition but also identification of the speaker's gender, age, emotional, and social status. Our purpose in this study is to evaluate whether temporal cues are sufficient to support both speech and speaker recognition. Ten cochlear-implant and six normal-hearing subjects were presented with vowel tokens spoken by three men, three women, two boys, and two girls. In one condition, the subject was asked to recognize the vowel. In the other condition, the subject was asked to identify the speaker. Extensive training was provided for the speaker recognition task. Normal-hearing subjects achieved nearly perfect performance in both tasks. Cochlear-implant subjects achieved good performance in vowel recognition but poor performance in speaker recognition. The level of the cochlear implant performance was functionally equivalent to normal performance with eight spectral bands for vowel recognition but only to one band for speaker recognition. These results show a disassociation between speech and speaker recognition with primarily temporal cues, highlighting the limitation of current speech processing strategies in cochlear implants. Several methods, including explicit encoding of fundamental frequency and frequency modulation, are proposed to improve speaker recognition for current cochlear implant users. © *2005 Acoustical Society of America.* [DOI: 10.1121/1.1944507]

PACS number(s): 43.71.Bp, 43.71.Es, 43.72.Fx, 43.66.Fe [KWG] Pages: 1055–1061

I. INTRODUCTION

Natural speech utterances carry information not only about what is being said but also about who says it (e.g., gender, age, ethnicity, and emotional state). Acoustic cues encoding "what" and "who" are widely distributed in both gross and fine spectral and temporal domains and can be influenced by physiological, behavioral, and cultural factors (Ladefoged and Broadbent, 1958; Stevens *et al.*, 1968; Atal, 1972; Johnson *et al.*, 1984; Childers and Wu, 1991; Wu and Childers, 1991; Stevens, 2002). For example, spectral peaks or formant frequencies that are critical for speech recognition also carry information regarding a speaker's identity as they reflect the individual speaker's anatomical and physical properties such as vocal tract size, shape and position (Fellowes *et al.*, 1997; Remez *et al.*, 1997). Conversely, temporal waveform periodicity or fundamental frequency (F0) that is typically correlated with a speaker's gender can influence speech recognition (Whalen *et al.*, 1993; Holt *et al.*, 2001) or directly carry lexical information in tonal languages (Liang, 1963).

While traditional research has focused on spectral cues, the temporal waveform envelope has been extensively studied in speech recognition (Van Tasell *et al.*, 1987; Rosen, 1992; Shannon *et al.*, 1995). It has been found that, in both real and simulated cochlear implant implementation, high levels of speech intelligibility can be achieved by encoding relatively slowly varying temporal envelopes that are extracted from one to several numbers of frequency bands (Wilson *et al.*, 1991; Dorman and Loizou, 1998; Zeng *et al.*, 2002). Recently, the utility of the temporal envelope cue has been extended to Mandarin tone recognition (Fu *et al.*, 1998; Xu and Pfingst, 2003; Zeng *et al.*, 2005) as well as other aspects of spoken language processing such as speaker identification (Cleary and Pisoni, 2002; Kong *et al.*, 2003; McDonald *et al.*, 2003; Fu *et al.*, 2004; Gonzalez and Oliver, 2005).

Cleary and Pisoni (2002) tested the effect of linguistic content (fixed sentence versus varied sentence) on talker discrimination between two females in 44 school-aged deaf children who had used the cochlear implant for at least 4 years. They found that these children achieved significantly higher than chance (50%) performance (mean percent correct score =68%) when the sentence was fixed, but produced essentially chance level performance at 57% correct when the sentence was varied. Their results suggest that the cochlear-implant users could not reliably recognize an unfamiliar talker's voice when the linguistic content varied. McDonald *et al.* (2003) replicated this finding using word stimuli in 21 adult cochlear-implant users and 24 normal-hearing listeners who listened to processed stimuli simulating the Nucleus SPEAK strategy (6 of 20 channel peaking). They found a similar linguistic effect on talker discrimination by both groups of subjects.

[a)]Portions of this work were presented at the 26th Midwinter Meeting of the Association for Research in Otolaryngology, Daytona Beach, Florida, 2003.

[b)]Current address: USC School of Dentistry, Los Angles, California 90089.

[c)]Corresponding author: 364 Med Surge II, University of California, Irvine, California 92697-1275. Telephone: (949) 824-1539; fax: (949) 824-5907; electronic mail: fzeng@uci.edu

Fu *et al.* (2004) used vowel materials to test gender discrimination in 11 adult cochlear-implant users and found a great deal of variability in performance ranging from 70% to 95% correct. They also performed the same task in a group of normal-hearing listeners and varied systematically the number of spectral bands and the temporal envelope cutoff frequencies. The result showed that the implant users produced performance equivalent to normal performance with four to eight spectral bands. Most interestingly, they found a contrast between speaker and vowel recognition in the four-band condition: Speaker identification was significantly improved as a function of the temporal envelope frequency from 20 to 320 Hz but vowel recognition did not under the same condition.

Gonzalez and Oliver (2005) also examined both gender and speaker identification as a function of the number of spectral bands in normal-hearing listeners. They used Spanish sentence materials and processed them using either sinusoidal and noise carriers. Similar to the findings of Fu *et al.*, Gonzalez and Oliver found that gender and speaker identification is systematically improved with the number of bands. In addition, they found a surprising result that the sinusoidal carrier produced significantly better performance than the noise carrier, particularly when the number of spectral bands was small. Previous studies on speech recognition in quiet found no such carrier effect (Dorman *et al.*, 1997), although recent studies on speech recognition in noise have hinted at a similar carrier effect (Nie *et al.*, 2003). One interpretation of this surprising carrier effect on speaker identification is that the temporal envelope cue, particularly the fundamental frequency, is better encoded by the sinusoidal carrier than the noise carrier (Gonzalez and Oliver, 2005). Another interpretation is that the sinusoidal carrier produces better modulation detection than the noise carrier and possibly resolved sidebands, particularly at low frequencies, to allow the normal-hearing listeners to directly hear out the voice pitch cue (Kohlrausch *et al.*, 2000; Zeng, 2003).

The above-mentioned studies implicated strongly that current cochlear implant users do not receive sufficient acoustic cues to support speaker identification and underscored the importance of extracting and encoding the temporal fine structure in cochlear implants (Oppenheim and Lim, 1981; Smith *et al.*, 2002; Nie *et al.*, 2005). Oppenheim and Lim (1981) independently manipulated Fourier amplitude and phase spectra and sometimes mixed one stimulus's amplitude spectrum with another stimulus's phase spectrum to demonstrate that phase provides critical information for auditory and visual perception. However, the importance of phase information has been largely ignored in the implant field until the Smith *et al.* chimera experiment (Smith *et al.*, 2002). Smith *et al.* mixed up one sound's temporal envelope with another sound's temporal fine structure to demonstrate their independent contributions to speech recognition and pitch perception. To overcome the difficulty of encoding the relatively rapid-varying temporal fine structure, Nie *et al.* (2005) derived slowly varying frequency modulations around the center frequency of a particular subband and found them to be effective in separating one speaker from another to achieve better speech recognition in noise.

Our goal for the present study was twofold. The first goal was to delineate the relative contributions of temporal envelope and fine structure to speech and speaker recognition. The second goal was to identify novel coding strategies to improve speaker identification in cochlear-implant users. To achieve these goals, the present study used the same vowel stimuli to collect systematically both vowel and speaker identification in normal-hearing and cochlear-implant subjects. The normal-hearing subjects listened to vowel syllables (in /hVd/ context) from ten speakers. These syllables included both the original unprocessed stimuli and the processed stimuli to contain either the temporal envelope cue or additionally the slowly varying frequency modulation cue. Performance for the processed stimuli was measured as a function of the number of frequency bands from 1 to 32. The cochlear-implant subjects performed the same task, but with only the original unprocessed stimuli. As a control, vowel recognition was also measured using identical stimuli from the same ten speakers in both normal-hearing and cochlear-implant subjects.

II. METHODS

A. Subjects

Six normal-hearing adults between the ages of 18 and 32 years and ten post-lingually deafened implant users between the ages of 49 and 74 years participated in the experiments. The implant subjects included 1 Ineraid device user (with a Med El CIS speech processor), 6 Nucleus users (with 3 SPEAK and 3 ACE users), and 3 Clarion users (with 1 CIS and 2 PSP users). Each implant subject had used the device for at least one year at the time of test. All participants were native English speakers. Additional demographic information can be found in Table I.

B. Stimuli

Stimuli consisted of 12 vowel tokens in the /hVd/ context and were originally recorded and analyzed by Hillenbrand and his colleagues (Hillenbrand *et al.*, 1995). Instead of the traditionally used sentence materials, the vowel stimuli were chosen because they could be used repetitively for the large number of experimental conditions employed in the present study, and additionally they were generally free of linguistic and speech rate/rhythm cues. In the speaker identification experiment, only two sets of three vowels were selected. One set (/had/, /heed/, and /hawd/) was used for practice and training purpose while the other set (/herd/, /hid/, and /hoed/) was used for the experiment. These tokens were chosen to ensure each set had high/high, high/low, and low/high F1/F2 values. Ten speakers including three men, three women, two boys, and two girls were used to form a total of 60 tokens. In the vowel recognition experiment, all 12 vowels were used. The same ten speakers produced a total of 120 tokens that were used for both practice and experiment purposes.

The original Hillenbrand stimuli were first pre-emphasized by a first-order high-pass Butterworth filter at 1200 Hz. The pre-emphasized stimuli were then bandpassed using fourth-order elliptic bandpass filters to produce 1, 4, 8,

TABLE I. Biographical data on cochlear implant subjects.

Subject #	Gender	Age (years)	Age of loss	Year of implantation	Etiology	Device	Strategy
1	M	71	39	1978	Meniere's	Ineraid	CIS
2	M	62	40	1990	Trauma	Nucleus 22	SPEAK
3	M	62	45	1995	Genetic	Nucleus 22	SPEAK
4	F	70	30	1998	Otosclerosis	Nucleus 22	SPEAK
5	F	49	9	1999	Unknown	Nucleus 24	ACE
6	F	69	44	1997	Virus	Nucleus 24	ACE
7	F	70	30	2000	High fever	Nucleus 24	ACE
8	F	68	34	1998	Autoimmune	Clarion I	CIS
9	F	72	46	2001	Nerve	Clarion II	PSP
10	F	74	57	2000	SNHL	Clarion II	PSP

16, and 32 subbands (Greenwood, 1990). The temporal envelope was extracted from each sub-band by full-wave rectification and low-pass filtering with a 500 Hz cutoff frequency. The slowly varying frequency modulation was extracted from each sub-band using a pair of phase-orthogonal demodulators with a cosine and sine carrier at the center frequency of each sub-band (Nie *et al.*, 2005). The frequency modulation had a bandwidth of 500 Hz and a modulation rate at 400 Hz. The frequency modulation extracted and preserved both the within-band and the cross-band phase information.

To produce stimuli with primarily temporal cues, the band-specific envelope was used to amplitude modulate a fixed carrier whose frequency was equal to the sub-band center frequency. To produce stimuli containing the slowly varying frequency modulation, the phase component was first recovered by integration of the frequency modulations before applying amplitude modulation by the temporal envelope (Nie *et al.*, 2005). Finally, before the summation of the sub-band signals, the same bandpass filter as the analysis bandpass filter was applied to both AM and AM+FM sub-band stimuli. This bandpass filter would effectively remove spectral differences between AM and AM+FM stimuli.

C. Procedure

Computer interfaces using MATLAB were developed for both speaker and vowel recognition experiments. Push buttons displayed on a computer monitor were created to correspond to a closed set of choices. For the speaker identification interface, ten push buttons were displayed in two rows. Row 1 corresponded to Male 1, Boy 1, Male 2, Boy 2, Male 3; Row 2 corresponded to Female 1, Girl 1, Female 2, Girl 2, Female 3. For the vowel recognition experiment, 12 corresponding pushbuttons were displayed on the interface.

Experiments were conducted in a double-walled sound treated booth (IAC). Stimuli were presented at 65 dBA via either a Sennheiser headset (HDA200) monaurally to normal-hearing listeners or a TANNOY Reveal speaker to cochlear-implant listeners. In the speaker identification experiment, all subjects received one to two hour training by systematically and/or randomly selecting push buttons to listen to the corresponding speaker's voice. All subjects underwent five practice rounds before formal data collection for the experimental condition. Each practice round consisted of 60 stimuli, including 2 presentations of 1 set of 3 vowels from 10 speakers. During testing, stimuli were presented randomly and the subject was subsequently asked to choose the correct speaker. Feedback was given after each selection by indicating whether the subject's choice was correct or incorrect via highlighting the push button corresponding to the correct answer.

After five practice rounds, the experimental test was conducted using the other set of three vowels to which the subject had not been exposed. In the vowel recognition experiment, the same procedure as in the speaker identification experiment was used except for a different task (recognizing 12 vowels instead of 10 speakers), and a different user interface. Only one practice session was conducted.

III. RESULTS

A. Original stimuli

Figure 1 shows training data from 5 practice sessions, as well as the test session (#6) for both normal-hearing (tri-

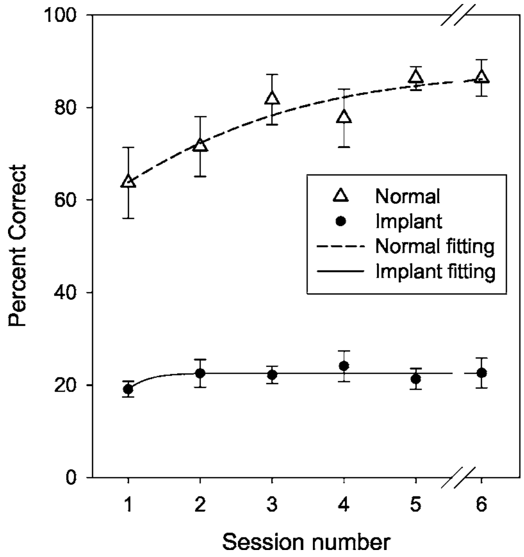

FIG. 1. Learning curve for speaker recognition in normal-hearing (open triangles) and cochlear-implant (filled circles) subjects. Sessions #1–5 represent the training period while session #6 represents the test run. Error bars represent plus and minus one standard error. The lines represent fitting of the learning curve with a sigmoidal function.

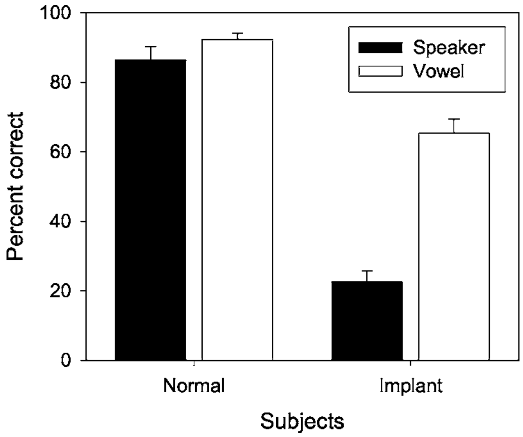

FIG. 2. Average performance for speaker (filled bars) and vowel (open bars) recognition in normal-hearing and cochlear-implant subjects. Error bars represent plus and minus one standard error. The chance performance is 10% for speaker recognition and 8% for vowel recognition.

angles) and cochlear-implant (circles) subjects. The normal subjects showed a significant learning effect with average performance increasing from 64% correct in session 1 to a plateau at about 84% in session 3 (paired *t* test, $p < 0.01$). In contrast, the implant subjects performed significantly more poorly than normal-hearing subjects with a plateau at approximately a 20% correct level. In addition, the three-percentage point training effect between sessions one and six was not significant ($p > 0.1$). A sigmoid function was used to fit the training data, showing an asymptotic performance of 88% and 23% correct for normal and implant subjects, respectively. Finally, there was no significant difference ($p > 0.1$) between the last practice run (session five) and the test run (session six) for both normal and implant subjects.

Figure 2 contrasts the overall performance between speaker (filled bars) and vowel (open bars) recognition in both normal and implant users. ANOVA with a between-subjects, faxed-factor design revealed a highly significant main effect for both the subjects $[F(1,28) = 165.1, p < 0.01]$ and the tasks $[F(1,28) = 47.8, p < 0.01]$. The normal subjects performed significantly better than the implant subjects on both tasks, with 86% correct for speaker recognition and 92% correct for vowel recognition, as opposed to 23% correct for speaker recognition and 65% correct for vowel recognition in implant subjects. The difference between speaker and vowel recognition was insignificant in normal subjects ($p > 0.05$) but was significant in cochlear-implant subjects ($p < 0.05$).

Figure 3 shows individual data from the ten implant subjects to further highlight the difference between speaker and vowel recognition. The individual score increases from 10% (chance performance) to 43% correct for speaker recognition. Had there been a strong correlation between the two tasks, a similar increasing trend would be observed for the individual performance for vowel recognition. Instead, an insignificant correlation was found ($r = 0.37, p > 0.05$), accounting for only 14% variability in the data.

B. Processed stimuli

Figure 4 compares the performance in both speaker (left panel) and vowel (right panel) recognition as a function of the number of spectral bands. The performance with the temporal envelope cue is represented by filled triangles (AM) while the performance with the additional frequency modulation cue is represented by open circles (AM+FM). For comparison, the cochlear implant performance is also included as the hatched bar, with its height corresponding to the mean score and its position on the *x* axis indicating the equivalent number of the AM bands.

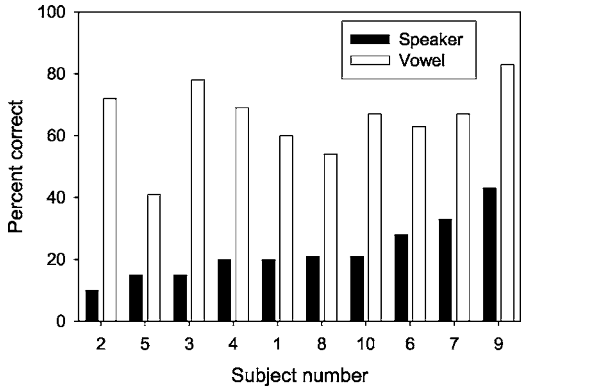

FIG. 3. Individual performance for speaker (filled bars) and vowel (open bars) recognition in cochlear-implant subjects. Individual data are ranked by the speaker recognition performance with the subject number corresponding to that in Table I.

In the speaker recognition task, the overall performance for the AM only condition was increased from 28% correct with one band to 76% correct with 32 bands. The corresponding performance for the AM+FM condition was from 46% to 82% correct. A repeated ANOVA shows a significant effect for both the processing [AM vs AM+FM: $F(1,18) = 564.7$, $p < 0.01$] and the number of bands $[F(4,18) = 504.2, p < 0.01]$. With four bands, the AM+FM condition produced the greatest improvement of 36 percentage points over the AM condition. Even with 32 bands, the AM+FM condition still resulted in significantly better performance than the AM condition by 6 percentage points ($p < 0.05$).

The vowel recognition performance was similar to the

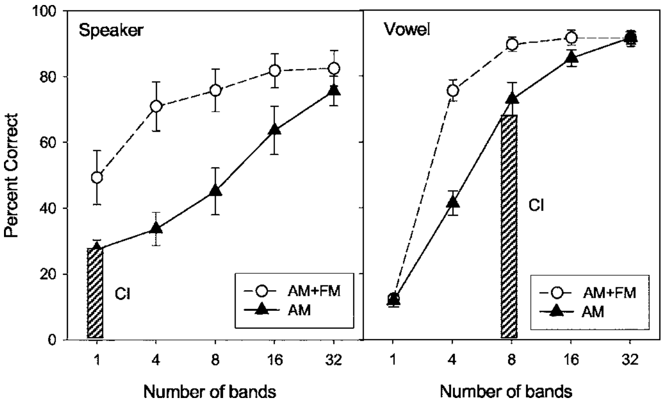

FIG. 4. Average performance for speaker (left panel) and vowel (right panel) recognition as a function of spectral bands in normal-hearing subjects. Filled triangles represent data obtained with the amplitude modulation cue (AM) while open circles represent data with both the amplitude and frequency modulation cues (AM+FM). Cochlear-implant performance is represented by the hatched bar, with its height corresponding to the mean score and its position on the *x* axis, indicating the equivalent number of the AM bands.

TABLE II. Stimulus-response or confusion matrix for speaker identification in cochlear-implant subjects.

	Man 1	Boy 1	Man 2	Boy 2	Man 3	Woman 1	Girl 1	Woman 2	Girl 2	Woman 3
Man 1	**9**	7	4	6	1	4	2	9	2	6
Boy 1	4	**8**	2	4	0	6	9	5	6	4
Man 2	4	2	**20**	1	16	0	1	3	0	1
Boy 2	7	1	4	**6**	1	5	5	8	6	5
Man 3	7	2	6	0	**32**	0	0	1	0	0
Woman 1	7	4	3	5	0	**7**	5	10	4	3
Girl 2	2	4	2	9	0	4	**11**	10	4	3
Woman 2	3	9	2	7	0	4	8	**10**	7	2
Girl 2	4	6	1	7	0	4	6	7	**7**	6
Woman 3	5	4	3	7	0	10	8	3	6	**3**

speaker recognition performance. However, several apparent differences were present, including a significant interaction between the processing and the task $[F(1,18)=35.6, p<0.01]$, and between the number of bands and the task $[F(1,18)=107.8, p<0.01]$. To demonstrate this interaction, first note the one-band results showing significantly better performance for speaker recognition than for vowel recognition for both the AM and the AM+FM conditions ($p<0.05$). Second, note the insignificant difference in performance between the speaker and vowel recognition with four bands ($p>0.05$). Third, note the reversed pattern showing better vowel recognition than speaker recognition with eight and more bands.

Notice, finally, that the equivalent number of bands for the cochlear-implant subjects is highly dependent on the task. In the speaker recognition task, the implant subjects performed at a level that was equivalent to the performance achieved by normal subjects with only one band. In contrast, the implant subjects were able to achieve a high level of performance on the vowel recognition task that was equivalent to eight bands for the normal subjects.

IV. DISCUSSION

A. Speaker versus vowel recognition

The most striking finding in the present study is the apparent disassociation of the use of temporal envelope cues between speaker and vowel recognition. This disassociation can best be observed by poor performance for speaker recognition but good performance for vowel recognition in the cochlear-implant subjects (Fig. 2). This disassociation is further enhanced by a lack of correlation between speaker and vowel recognition in the implant subjects (Fig. 3). In cochlear implant simulation, this disassociation is best illustrated by the interaction between the number of bands and the listening tasks (Fig. 4). With only 1-band envelope, speaker recognition was 16 percentage points significantly better than vowel recognition; but with 8-band envelopes, speaker recognition was 28 percentage points significantly worse. The disassociation results suggest that depending on the availability of acoustic cues, the brain may use different strategies to process information regarding speaker and vowel recognition.

The disassociation results also suggest that speaker and speech (vowel) recognition may place different weights on different acoustic cues. Speaker recognition relies more on low-frequency cues that can be derived from temporal envelopes, while vowel recognition relies more on high-frequency spectral cues that require a large number of bands. This suggestion is consistent with the traditional view that acoustic cues carrying speaker information are highly related to fundamental frequency and that acoustic cues carrying speech information are highly related to formant frequencies, particularly the second formant frequency (French and Steinberg, 1947).

B. Analysis of error patterns

The speaker pool used in the present study included both gender and age factors. Although both actual and simulated implant performance was low for speaker recognition, it was still possible that the implant subjects could identify gender and age, but were only confused within categories. To analyze the error patterns in speaker recognition, classic sequential information transfer analysis (SINFA) was performed (Wang and Bilger, 1973).

Table II shows the confusion matrix for speaker recognition in eight of the ten cochlear-implant subjects. Subject #1 and #2 were not available as their data were collected before the information regarding the speaker confusion pattern was recorded in the program. The stimuli were represented as rows while the responses were represented as columns. A total of 488 tokens were pooled from 8 subjects with each contributing to 61 responses (10 speakers × 3 tokens × 2 presentation + 1 randomly selected token). The overall score was 23.2% correct, indistinguishable from the 22.6% score obtained from the 10 implant subjects.

SINFA (Wang and Bilger, 1973) shows that the cochlear-implant subjects were only able to receive 4.7% gender information and 2.5% age information, corresponding to 62.7% and 60.9% overall percent correct, respectively. The percent information transmitted improved slightly to 9.3% for gender discrimination when age had been accounted for and to 4.3% for age discrimination when the gender had been accounted for. SINFA was also used to analyze the error patterns with one- and four-band conditions in normal-hearing subjects and found generally similar results to the implant pattern. Together, the present analysis of error patterns demonstrated that temporal envelope cues do not provide reliable information for speakers either across or within both gender and age categories.

C. Implications for cochlear implants

The present results highlight the limitation of current speech processing strategies in cochlear implants. Except for the infrequently used analog strategies, only temporal envelope information from several bands is extracted while the temporal fine structure information is discarded in the process. Given the limited number of functional channels available in current cochlear implants, it is clear that the temporal envelope cue is not sufficient to support reliable speaker recognition.

There are at least three ways to redesign current speech processing strategies to improve speaker recognition performance in cochlear-implant users. One way is to explicitly encode the fundamental frequency information. In an earlier but now abandoned speech processing strategy (Skinner *et al.*, 1991), information regarding the fundamental frequency along with the first and/or second formant frequencies was extracted and delivered to the cochlear implant via pulsatile stimulation patterns following the changes in fundamental frequency. To our knowledge, no study had been performed to directly evaluate this earlier strategy's performance in speaker recognition. It is possible that the fundamental frequency information can be reintroduced as a carrier frequency in the modern speech processing strategies. A simulation of such a processing strategy has been shown to improve Mandarin tonal recognition (Lan *et al.*, 2004).

Another way to improve upon the current cochlear implants is to extract the slowly varying frequency modulation and deliver it to cochlear implants (Nie *et al.*, 2005). The slowly varying frequency modulation does not explicitly extract fundamental frequency but does contain information regarding the direction and rate of both fundamental and formant movements. The present simulation result shows that this slowly varying frequency modulation could produce significantly better performance in speaker recognition, even with one or four bands.

A third way to improve upon the current cochlear implants is to introduce a high-frequency or noise conditioner to improve frequency representation in the temporal envelope domain at the auditory nerve level (Rubinstein, 1995; Morse and Evans, 1996; Litvak *et al.*, 2003). The hope is that the conditioner would improve frequency discrimination (Zeng *et al.*, 2000), which would in turn improve speaker identification based on relatively low frequencies in the envelope domain. While it is not clear which exact strategy or a combination of strategies might work, it is clear from the present data that current speech processing strategies need to be changed to improve speaker recognition performance in cochlear implant users.

V. CONCLUSIONS

Speaker and vowel recognition performance was measured in ten cochlear-implant and six normal-hearing subjects. The speakers included three men, three women, two boys, and two girls. The stimuli were 12 vowels in /hVd/ context from the Hillenbrand study (Hillenbrand *et al.*, 1995). The main findings are as follows.

(1) Current cochlear-implant users are able to achieve good performance in vowel recognition (65% correct) but poor performance in speaker recognition (23% correct).
(2) Implant performance is functionally equivalent to normal performance of eight spectral bands with temporal envelopes for vowel recognition but only one band for speaker recognition.
(3) A slowly varying form of frequency modulation can improve significantly the speaker recognition performance and should be encoded in future cochlear implants.
(4) The present result supports the hypothesis that speaker and speech recognition with primarily temporal cues involves two independent processes.

ACKNOWLEDGMENTS

We thank Dr. Kaibao Nie for providing the MATLAB program to process the stimuli, University College London for the SINFA program, and Abby Copeland, Molly Pitassi, Ken Grant and two anonymous reviewers for their helpful comments on the manuscript. The work was supported by the National Institutes of Health (2RO1DC02267) and UCI Undergraduate Research Opportunity Program (UROP to Vongphoe who was also selected as Researcher of the Month in September, 2002).

Atal, B. S. (**1972**). "Automatic speaker recognition based on pitch contours," J. Acoust. Soc. Am. **52**, 1687–1697.

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.

Dorman, M. F., and Loizou, P. C. (**1998**). "The identification of consonants and vowels by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normal-hearing subjects using simulations of processors with two to nine channels," Ear Hear. **19**, 162–166.

Dorman, M. F., Loizou, P. C., 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. M., Remez, R. E., and Rubin, P. E. (**1997**). "Perceiving the sex and identity of a talker without natural vocal timbre," Percept. Psychophys. **59**, 839–849.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.

Fu, Q. J., Chinchilla, S., and Galvin, 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., Zeng, F. G., Shannon, R. V., and Soli, S. D. (**1998**). "Importance of tonal envelope cues in Chinese speech recognition," J. Acoust. Soc. Am. **104**, 505–510.

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. (in press).

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

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

Holt, L. L., Lotto, A. J., and Kluender, K. R. (**2001**). "Influence of fundamental frequency on stop-consonant voicing perception: A case of learned covariation or auditory enhancement?" J. Acoust. Soc. Am. **109**, 764–774.

Johnson, C. C., Hollien, H., and Hicks, J. W. (**1984**). "Speaker identification utilizing selected temporal speech features," J. Phonetics **36**, 93–100.

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 modulation and frequency modulation to auditory perception: II. Melody, tone and speaker identification," Abstract of the 26th Annual Midwinter Research Meeting, Vol. 26, pp. 213–214.

Ladefoged, P., and Broadbent, D. E. (**1958**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98–104.

Lan, N., Nie, K. B., Gao, S. K., and Zeng, F. G. (**2004**). "A novel speech-processing strategy incorporating tonal information for cochlear implants," IEEE Trans. Biomed. Eng. **51**, 752–760.

Liang, Z. A. (**1963**). "Auditory perceptual cues in Mandarin tones," Acta Phys. Sin. **26**, 85–91.

Litvak, L. M., Delgutte, B., and Eddington, D. K. (**2003**). "Improved temporal coding of sinusoids in electric stimulation of the auditory nerve using desynchronizing pulse trains," J. Acoust. Soc. Am. **114**, 2079–2098.

McDonald, C. J., Kirk, K. I., Krueger, T., Houston, D., and Sprunger, A., (**2003**). "Talker discrimination by adults with cochlear implants," *Abstracts of the 26th Annual Midwinter Research Meeting of the Association for Research in Otolaryngology*, Daytona Beach, Florida.

Morse, R. P., and Evans, E. F. (**1996**). "Enhancement of vowel coding for cochlear implants by addition of noise," Nat. Med. **2**, 928–932.

Nie, K., Stickney, G., and Zeng, F. G. (**2003**). "Independent contributions of amplitude modulation and frequency modulation to auditory perception: I. Consonant, vowel and sentence recognition," *Abstracts of the 26th Annual Midwinter Research Meeting of the Association for Research in Otolaryngology*, Daytona Beach, Florida.

Nie, K., Stickney, G., and Zeng, F. G. (**2005**). "Encoding fine structure to improve cochlear implant performance in noise," IEEE Trans. Biomed. Eng. **52**, 64–73.

Oppenheim, A. V., and Lim, J. S. (**1981**). "The importance of phase in signals," Proc. IEEE **69**, 529–541.

Remez, R. E., Fellowes, J. M., and Rubin, P. E. (**1997**). "Talker identification based on phonetic information," J. Exp. Psychol. Hum. Percept. Perform. **23**, 651–666.

Rosen, S. (**1992**). "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. London, Ser. B **336**, 367–373.

Rubinstein, J. T. (**1995**). "Threshold fluctuations in an *N* sodium channel model of the node of Ranvier," Biophys. J. **68**, 779–785.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Skinner, M. W., Holden, L. K., Holden, T. A., Dowell, R. C., Seligman, P. M., Brimacombe, J. A., and Beiter, A. L. (**1991**). "Performance of post-linguistically deaf adults with the Wearable Speech Processor (WSP III) and Mini Speech Processor (MSP) of the Nucleus Multi-Electrode Cochlear Implant," Ear Hear. **12**, 3–22.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nature (London) **416**, 87–90.

Stevens, K. N. (**2002**). "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am. **111**, 1872–1891.

Stevens, K. N., Williams, C. E., Carbonell, J. R., and Woods, B. (**1968**). "Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material," J. Acoust. Soc. Am. **44**, 1596–1607.

Van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. P. (**1987**). "Speech waveform envelope cues for consonant recognition," J. Acoust. Soc. Am. **82**, 1152–1161.

Wang, M. D., and Bilger, R. C. (**1973**). "Consonant confusions in noise: A study of perceptual features," J. Acoust. Soc. Am. **54**, 1248–1266.

Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (**1993**). "FO gives voicing information even with unambiguous voice onset times," J. Acoust. Soc. Am. **93**, 2152–2159.

Wilson, B. S., Finley, C. C., Lawson, D. T., Wolford, R. D., Eddington, D. K., and Rabinowitz, W. M., (**1991**). "Better speech recognition with cochlear implants," Nature (London) **352**, 236–238.

Wu, K., and Childers, D. G. (**1991**). "Gender recognition from speech. Part I: Coarse analysis," J. Acoust. Soc. Am. **90**, 1828–1840.

Xu, L., and Pfingst, B. E. (**2003**). "Relative importance of temporal envelope and fine structure in lexical-tone perception," J. Acoust. Soc. Am. **114**, 3024–3027.

Zeng, F. G. (**2003**). "Compression and cochlear implants," *Compression: From Cochlea to Cochlear Implants*, edited by S. P. Bacon, A. N. Popper, and R. R. Fay (Springer-Verlag, New York), Vol. 17, pp. 184–220.

Zeng, F. G., Fu, Q. J., and Morse, R. (**2000**). "Human hearing enhanced by noise," Brain Res. **869**, 251–255.

Zeng, F. G., Grant, G., Niparko, J., Galvin, J., Shannon, R., Opie, J., and Segel, P. (**2002**). "Speech dynamic range and its effect on cochlear implant performance," J. Acoust. Soc. Am. **111**, 377–386.

Zeng, F. G., Nie, K., Stickney, G. S., Kong, Y. Y., Vongphoe, M., Bhargave, A., Wei, C. G., and Cao, K. (**2005**). "Speech recognition with amplitude and frequency modulations," Proc. Nat. Acad. Sci. USA **102**, 2293–2298.