Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants

Andrew Faulkner,[a] Stuart Rosen, and Clare Smith
Department of Phonetics and Linguistics, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, United Kingdom

(Received 8 April 2000; accepted for publication 24 July 2000)

Recent simulations of continuous interleaved sampling (CIS) cochlear implant speech processors have used acoustic stimulation that provides only weak cues to pitch, periodicity, and aperiodicity, although these are regarded as important perceptual factors of speech. Four-channel vocoders simulating CIS processors have been constructed, in which the salience of speech-derived periodicity and pitch information was manipulated. The highest salience of pitch and periodicity was provided by an explicit encoding, using a pulse carrier following fundamental frequency for voiced speech, and a noise carrier during voiceless speech. Other processors included noise-excited vocoders with envelope cutoff frequencies of 32 and 400 Hz. The use of a pulse carrier following fundamental frequency gave substantially higher performance in identification of frequency glides than did vocoders using envelope-modulated noise carriers. The perception of consonant voicing information was improved by processors that preserved periodicity, and connected discourse tracking rates were slightly faster with noise carriers modulated by envelopes with a cutoff frequency of 400 Hz compared to 32 Hz. However, consonant and vowel identification, sentence intelligibility, and connected discourse tracking rates were generally similar through all of the processors. For these speech tasks, pitch and periodicity beyond the weak information available from 400 Hz envelope-modulated noise did not contribute substantially to performance. © *2000 Acoustical Society of America.* [S0001-4966(00)05810-0]

PACS numbers: 43.71.Ky, 43.71.Bp, 43.66.Ts [CWT]

I. INTRODUCTION

Pitch variation, and the presence of periodic and/or aperiodic excitation, are widely held to be important cues for the perception of speech. However, surprisingly little is known of their contribution to speech intelligibility except in what may be a special case, that of auditory signals that contain no spectral structure. With such signals, these factors contribute in several important ways. The timing of periodic and aperiodic excitation are dominant temporal cues to consonant identity (Faulkner and Rosen, 1999). Furthermore, for the audio-visual perception of connected speech, both voice pitch and the timing of voiced excitation provide distinct elements of complementary support to visual cues (Breeuwer and Plomp, 1986; Grant *et al.*, 1985; Risberg, 1974; Risberg and Lubker, 1978; Rosen, Fourcin, and Moore, 1981). Speech presented through a cochlear implant, or through a vocoder-like simulation of an implant speech processor, is represented by a relatively small number of spectral bands, each conveying temporal envelope information. It may be expected, then, that the temporal information that contributes to speech perception through such processing is similar to the temporal information that dominates perception from signals that convey no spectral information.

Vocoder-like speech-processing methods have been used in a number of recent studies that aim to simulate cochlear implant speech processors (Dorman, Loizou, and Rainey, 1997a, 1997b; Rosen, Faulkner, and Wilkinson, 1999; Shannon *et al.*, 1995; Shannon, Zeng, and Wygonski, 1998). These simulations represent the spectro-temporal information delivered to the auditory nerve by continuous interleaved sampling (CIS) processors (Wilson *et al.*, 1991). In a CIS implant, the signals presented along the electrode array represent amplitude envelopes extracted from a series of bandpass filters. These envelopes, typically smoothed to carry temporal information below 400 Hz, are imposed on biphasic pulse carriers that generally have a rate between 1 and 2 kHz.

The simulation studies performed so far have paid little attention to the nature of the temporal cues provided. Rather, the focus has been on the role of spectral resolution (Dorman, Loizou, and Rainey, 1997b; Shannon *et al.*, 1995) and the effects of shifts of the spectral envelope (Dorman *et al.*, 1997a; Rosen *et al.*, 1999; Shannon *et al.*, 1998). Here, we focus on the contributions to speech intelligibility that can be attributed to speech-related pitch information (i.e., variation in voice fundamental frequency) and to periodicity information (i.e., the presence of periodic laryngeal excitation or of aperiodic voiceless excitation).

Previous simulation studies have made use of either bandpass-filtered noise carriers, or a series of fixed-frequency sinusoidal carriers to deliver amplitude envelope information in selected frequency bands to the normal ear. Temporal cues to pitch variation, and to the simple presence of periodicity, are carried by the modulation of the pulse stimulation from a CIS processor as long as two conditions are met. The envelope smoothing filter must encompass the voice fundamental frequency range and the pulse stimulation

[a] Electronic mail: andyf@phon.ucl.ac.uk

rate must be sufficiently high to sample this frequency range adequately. Similarly, where vocoder simulations use sufficiently high envelope bandwidths to modulate noise carriers,[1] these too are capable of signaling pitch and periodicity for modulation rates up to a few hundred Hz (e.g., Pollack, 1969). However, the salience of the pitch of modulated noise is weak compared to that of harmonic sounds such as voiced speech, and it is important to establish the limitations that such simulations may have in respect to the transmission of pitch and periodicity. Little is known about the effects of the salience of periodicity in such simulations. Fu and Shannon (2000) report little effect of varying the envelope cutoff frequency between 16 and 400 Hz for English consonant materials with four-channel noise-excited vocoders. In Chinese, however, it has been shown that tonal cues carried by noise modulated by a 400-Hz bandwidth speech envelope can contribute to sentence-level speech perception using such simulations (Fu, Zeng, and Shannon, 1998).

A. Pitch and periodicity cues from a CIS cochlear implant processor

The representation of pitch variation and of speech periodicity for users of a CIS cochlear implant speech processor will depend not only on the extent to which the corresponding temporal information is contained in the extracted amplitude envelopes, but also on the extent to which the patient is able to process this information. This latter aspect is not well understood, although it is clear that there are very wide variations between patients. A study of periodic/aperiodic discrimination in single-channel implant users showed some patients to have good abilities in identifying periodic from aperiodic pulse stimulation, at least for stimuli of 200-ms duration (Fourcin *et al.*, 1979). However, except for one subject, the stimuli used in that study were directly periodic or aperiodic, not pulse carriers with periodic or aperiodic amplitude modulation. McDermott and McKay (1997) studied one individual implant patient under conditions comparable to CIS stimulation. Sinusoidal amplitude modulation of a 1200-Hz pulse train delivered to a single bipolar electrode pair allowed the discrimination of modulation rates differing by 3% to 4% around a 100-Hz rate. Around a 200-Hz rate, thresholds were between 4% and 27%, depending on the stimulation site. Other selected CIS implant processor users have also showed good ability in the pitch ranking of pulsatile stimulation that carries sinusoidal amplitude modulation up to modulation rates of 1 kHz (Wilson *et al.*, 1997). However, this last study gives rather limited information on pitch discrimination, since the ranked modulation rates differed in steps of 100 Hz.

B. Representation of pitch information in vocoder carriers

In normal hearing, pitch perception is thought to be based primarily on temporal cues derived from resolved lower-frequency harmonics, including the fundamental component, and also on periodicity cues in the temporal envelope in auditory filter channels driven by adjacent unresolved harmonics. Spectral details and overall spectral shape are also related to fundamental frequency, and are encoded by place within the limits of auditory frequency resolution. For quasi-periodic speech-like signals, a CIS implant processor would not be expected to deliver useful place-based spectral pitch cues within the voice fundamental frequency (Fx) range. The primary reason for this is that the channel bandpass filters are too wide to resolve individual harmonics of such fundamental frequencies.[2] In addition, the spectral shape of speech is constantly varying independently of fundamental frequency, so that spectral envelope is unlikely to be a reliable source of pitch information for speech. Hence, only envelope periodicity cues will be available to signal pitch for speech. The carrier in a CIS processor is a non-random high-rate pulse rather than the random noise typically used in simulations. For this reason, temporal modulation of the carrier related to Fx will be noise-free, and the neural responses to this stimulation are also likely to be strongly synchronized to the modulation (Wilson *et al.*, 1997).

This study introduces the use of frequency-controlled pulse carriers for voiced speech. Here, the carrier for voiced speech is a flat-spectrum pulse train whose period is controlled by voice fundamental frequency. The carrier is passed through a series of bandpass filters to control the frequency content of the different output bands. The use of such a carrier is not intended to represent the pulsatile stimulation of CIS, which cannot be accurately emulated in acoustic hearing. Rather, the intention is to achieve the highest possible pitch salience by providing a rich set of pitch cues both from individually resolved lower harmonics and from temporal envelope cues from the unresolvable higher harmonics. The noise carriers typical of most previous simulation studies necessarily lack harmonic content, and provide only temporal envelope cues to pitch. With noise carriers, the periodicity of the temporal envelope related to voice pitch will be noisy by virtue of the random nature of the carrier. Such random fluctuations in the carrier will be more significant in the lower vocoder bands, where the rate of the inherent envelope fluctuations of the filtered carrier is closer to the rate of the envelope fluctuations extracted from periodic speech. The random nature of noise carriers may well weaken pitch salience compared to that derived from CIS processors by those implant users who are able to fully process the temporal information carried by envelope-modulated pulse stimulation.

II. EXPERIMENTAL QUESTIONS

The studies reported here address two related issues, using a variety of segmental and connected speech tasks. Given that simulations of vocoder-like CIS speech processors deliver limited pitch and periodicity information, what impact does this have on speech intelligibility, and would more salient pitch and periodicity cues improve performance?

III. METHODS

A. Signal processing

Signal processing was implemented in real time, using the ALADDIN INTERACTIVE DSP WORKBENCH software (v1.02, AB Nyvalla DSP). It ran at a 16-kHz sample rate on a Loughborough Sound Images DSP card with a Texas Instru-

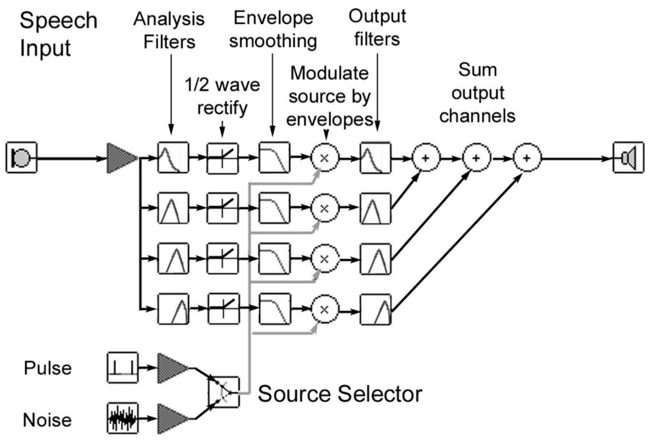

FIG. 1. Block diagram showing speech processing common to all processor simulations.

ments TMS320C31 processor. All processors used here had four channels, with the analysis and output filters being identical, so that the spectral representation was tonotopically accurate within the constraints of the limited spectral resolution. A block diagram of the common components of the processors is shown in Fig. 1. Each channel consisted of a series of blocks, comprising: a bandpass filter applied to the speech input; a rectifier and low-pass filter to extract the amplitude envelope from that spectral band; a multiplier that modulated a carrier signal by that envelope; a final bandpass filter, matching the analysis filter, shaped the spectrum of the modulated carrier signal.

The four analysis and output filter bands were based on equal basilar membrane distance (Greenwood, 1990). The filter slopes crossed at their −6-dB cutoff frequencies, these being 100, 392, 1005, 2294, and 5000 Hz. The bandpass analysis filters, and the corresponding output filters, were eighth-order elliptical IIR designs, with slopes in excess of 50 dB/octave, and stop bands at least 50 dB down on the passband. The amplitude envelope was extracted from each analysis filter output by half-wave rectification followed by a fourth-order elliptical low-pass filter, with a slope of about 48 dB/octave.

1. Speech-processing conditions

The various processing conditions are summarized in Table I. With one exception, the envelope extraction employed a 32-Hz low-pass filter, so that temporal information in the voice pitch and periodicity range was eliminated from the envelope. The salience of speech-derived pitch and periodicity was manipulated through the selection and control of the carrier signal. The fullest and most salient representation of pitch and periodicity was produced using processing similar to classic speech synthesizing vocoders (Dudley, 1939). Here, the carrier source during voiced speech was a pulse signal whose frequency followed that of the fundamental frequency of the speech input (Fx). The carrier source for voiceless speech was a random noise (symbolized as Nx). This condition is notated as $FxNx$. The pulse carrier was a monophasic pulse with a width of one sample (63 μS). Within the 8-kHz overall bandwidth of the processor, the spectral envelope of this pulse train and the noise source were both flat, and both source signals had the same rms level.

A processor similar to that used for condition $FxNx$ differed only in using a fixed 150-Hz pulse rate rather than a speech-derived pulse rate. This processor preserved the contrast between periodic and purely aperiodic excitation, while discarding voice pitch variation. It is designated as condition $VxNx$.

A third processor discarded both periodicity and pitch information and was produced by using a fixed-frequency 150-Hz pulse source for all speech input. This condition was designated *Mpulses* (monotone pulses).

Two further processors employed a filtered white-noise carrier for all speech. These are similar to the processors used by Shannon *et al.* (1995). They differed from each other only in the low-pass cutoff frequency of the envelope filters, which was either 400 Hz in condition *Noise400*, or 32 Hz in condition *Noise32*. The 400-Hz envelope cutoff was expected to allow speech periodicity and pitch information to be preserved in the extracted envelope. However, the perceptual salience of this information was not expected to be as high as for condition $FxNx$. The use of a 32-Hz cutoff frequency together with the 48-dB/octave slope of the envelope filter was expected to eliminate virtually all pitch and periodicity cues in condition *Noise32*.

With the exception of processor *Noise400*, all processors represent the spectral envelope of the input signal essentially identically. The spectral envelope signaled by processor *Noise400* differs in the representation of spectral envelope changes at a more rapid rate (up to 400 rather than 32 Hz). The spectra resulting from a pulsatile carrier inevitably differ from those from noise carriers in that harmonics

TABLE I. Summary of processor conditions (see the text for details).

Processor	Voiced speech carrier	Voiceless speech carrier	Envelope low-pass cutoff (Hz)	Expected salience of pitch and periodicity
Noise400	Noise	Noise	400	Both weak
Noise32	Noise	Noise	32	Both nil
$VxNx$	150-Hz pulse train	Noise	32	High for periodicity, nil for pitch
$FxNx$	Fx pulse train	Noise	32	Both high
Mpulses	150-Hz pulse train	150-Hz pulse train	32	Both nil

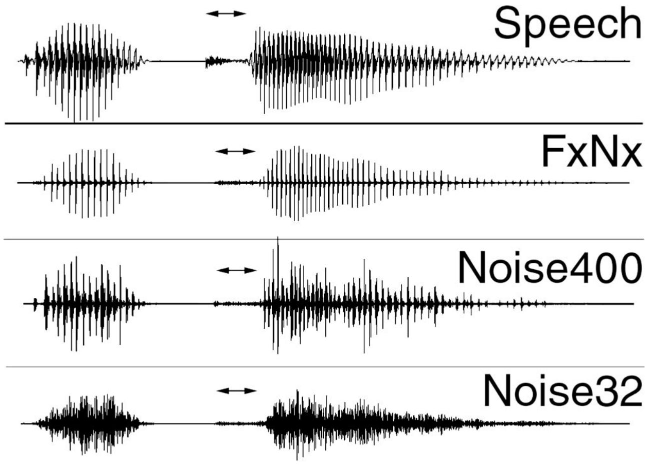

FIG. 2. Waveforms of speech input and output of band 3 from processors *FxNx*, *Noise400*, and *Noise32* for a male production of /ɑtɑ/. The arrows indicate the temporal extent of voiceless excitation in the input. The output of processor *VxNx* will differ from that of *FxNx* only in that, during voiced speech, the carrier-pulse rate is fixed at 150 Hz. The output from processor *Mpulses* will be similar to that from *Noise32*, except that the output is always periodic at a fixed rate of 150 Hz.

of the carrier are present. However, this spectral detail is unrelated to the spectral shape of the input signal. When the pulse rate is controlled by the speech fundamental frequency, this spectral detail is one source of pitch information.[3] Processors *Noise32* and *Mpulses* both eliminate temporal cues to the pitch and the periodicity of the input speech, and differ only in that the output is either always aperiodic or always periodic.

2. Voicing detection and source switching

All speech materials were accompanied by a laryngographic signal marking glottal closure. Before processing through the simulations, the raw laryngograph waveform was preprocessed to produce a single discrete pulse at each laryngeal closure. The processors took this pulse train as input in addition to the speech signal. A dc offset was added to the pulse input to ensure that it passed through zero, and a zero-crossing detector was employed to detect the pulse period. Alternate zero-crossings triggered the generation of a carrier pulse. A sample-and-hold with a 10-ms time constant was applied to the output of the zero-crossing detector and the output of this stage was used as a voicing detector. The voicing detector output, smoothed by a first-order 50-Hz low-pass filter, was used to switch between the pulse and a white-noise source. The input speech was delayed by 30 ms before the initial bandpass analysis filtering to allow accurate time alignment of the switching between the vocoder carrier signals with changing speech excitation.

B. Results of speech processing

Figure 2 shows the output of the third spectral channel of processors *FxNx*, *Noise400*, and *Noise32* for the intervocalic consonant /ɑtɑ/, together with the original speech. It illustrates the representation of fundamental frequency and periodicity in the various processed signals.

C. Speech perceptual tests

Auditory performance for segmental and connected-speech materials was measured using four standard procedures. The contributions of periodicity and pitch information conveyed by the different processors were measured by reference to performance with processor *Noise32*, which conveys neither periodicity nor pitch. With the exception of connected discourse tracking, no feedback was given.

1. Consonant identification

The consonant set contained 20 intervocalic consonants with the vowel /ɑ/. These comprised all the English consonants except for /ð,ʒ,h,ŋ/. Materials were from digital anechoic recordings presented at a 22.05-kHz sample rate and were from one female and one male talker, mixed in each test run. Both talkers had a standard Southern British English accent. Each run presented 40 consonants, with one consonant from each talker being selected at random from a set of six to ten tokens. Stimulus presentation was computer controlled. Subjects responded using the computer mouse to select one of 20 buttons on the computer screen that were orthographically labeled to represent each of the 20 consonants.

2. Vowel identification

17 b-vowel-d words from the same two talkers were used, again from digital anechoic recordings presented at a 22.05-kHz sample rate. Presentation was computer controlled. Each test run presented one token of each word from each of the two talkers, selected at random from a total set of six to ten tokens of each word from each talker. The vowel set contained ten monophthongs (in the words *bad, bard, bead, bed, bid, bird, bod, board, booed*, and *bud*) and seven diphthongs (in the words *bared, bayed, beard, bide, bode, boughed*, and *Boyd*). The spellings given here are those that appeared on the computer response buttons.

3. Sentence perception

BKB sentences from a different female talker with the same British accent were used, from an analog audio-visual recording on U-matic videotape (EPI Group, 1986; Foster *et al.*, 1993). Each test run used one list of 16 sentences with 50 scored key words per list.

4. Connected discourse tracking

Live voice connected discourse tracking (CDT: DeFilippo and Scott, 1978) was conducted by a third single female talker (author CS). In CDT, the talker wore laryngograph electrodes to provide a larynx period and voicing reference. Materials were taken from texts for students of English as a second language.

D. Pitch salience test

Pitch salience through each processor was examined by the use of tone glides. The stimuli were sawtooth waves, chosen as having a spectrum similar to that of voiced speech. Each was 500 ms in duration and had a linear fundamental frequency transition from start to end. Three fundamental frequency ranges were included, centered around 155, 220, and 310 Hz. The start and end frequencies of the glides varied in six steps from a ratio of 1:0.5 to a ratio of 1:0.93. The test was again presented under computer control. On each trial, subjects heard a single glide, and were asked to categorize it as either ''rising'' or ''falling'' in pitch. They responded by clicking on one of two response buttons labeled with a rising or falling line. No feedback was provided. Each single administration of this test presented one rising and one falling tone at each start-to-end frequency ratio in each of the three frequency ranges, with 36 stimuli in total.

IV. PROCEDURE

Five subjects, screened for normal hearing up to 4 kHz, were recruited for the consonant, vowel, sentence, and tone glide tests. For each of these tests, six testing blocks were presented, in which each of the four tests was administered once through each of the five processors. The first two blocks were treated as practice. Because only 21 BKB lists were available, one identical BKB list was presented repeatedly for the first two practice blocks. In the final four blocks, a different BKB list was presented on every occasion.

CDT was run subsequently, with two subjects who had taken part in the earlier tests and an additional four subjects who were also screened for normal hearing. The CDT testing used only four of the five processors, with the *VxNx* processor being excluded. Each of the total of six testing sessions included 10 min of CDT with each of the four processors used. Each of these 10-min blocks was scored in two subunits of 5-min duration. Unprocessed speech was presented at the start of the first session to familiarize subjects with the task and to estimate ceiling performance rates. The order of use of the four processors was counterbalanced in a different order for each subject over the six sessions.

V. RESULTS

A. Frequency glides

Psychometric functions for labeling of glide direction as a function of start-to-end frequency ratio are shown in Fig. 3. For processor *FxNx*, performance for all the glide stimuli was at very high levels, as would be expected given the direct representation of the signal frequency in the carrier signal. Even with the smallest start-to-end frequency ratio of 1:0.93, scores were around 90% correct. Both modulated-noise processors allowed a limited identification of the direction of pitch glides. Performance with processor *Noise400* was above 75% correct for ratios of 1:0.66 and larger. Performance with processor *Noise32* was poorer than with *Noise400*, but better than that shown by the fixed-frequency pulse processors *Mpulses* and *VxNx*. For these, performance was close to chance as would be expected. The above-chance performance with processors *VxNx* and *Mpulses* at the largest frequency ratios, and the somewhat higher level of performance with *Noise32*, can only be attributed to spectral envelope differences that arise as harmonics of the input signal shift between processor bands, since the 32-Hz envelope bandwidth of these processors cannot encode fundamental frequency.

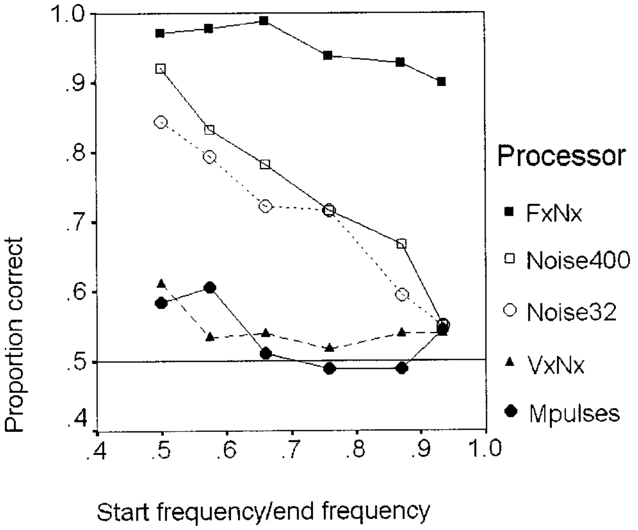

FIG. 3. Performance in labeling processed sawtooth wave frequency glides as a function of the ratio of the start and end frequencies (ignoring glide direction). Each point shows a mean score over four test sessions from 15 samples over the 3 frequency ranges and the 5 subjects. Chance performance is 50% correct.

Psychometric functions for the proportion of ''fall'' responses as a function of the log(base 10) of the start-to-end frequency ratio were estimated using a logistic regression applied to the group data. The resulting slope estimates and their 95%-confidence limits are shown in Fig. 4. The slope for processor *FxNx* is substantially steeper than that in all other conditions. The slope for the 400-Hz envelope bandwidth noise-carrier processor *Noise400* is slightly but significantly steeper than that for the *Noise32* processor. Slopes for the two fixed-period pulse processors *Mpulses* and *VxNx* are close to zero.

B. Vowel identification

Box and whisker plots of the group data with each processor are shown in Fig. 5. Scores were around 50% correct in all conditions. A repeated-measures analysis of variance

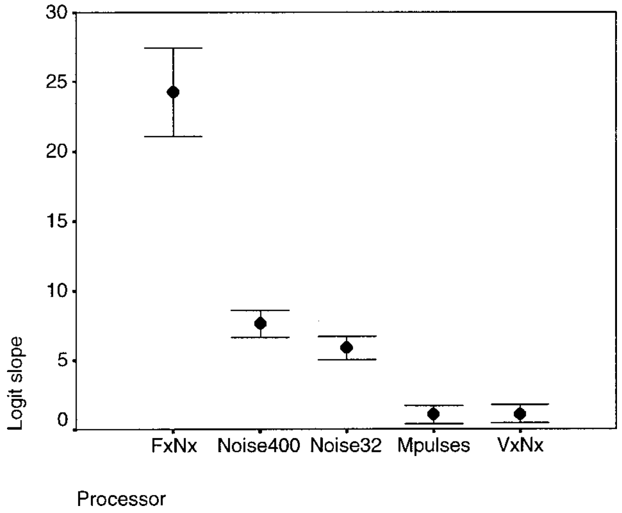

FIG. 4. Slopes of the psychometric functions estimated from a logistic regression of the proportion of ''fall'' responses as a function of the log(base 10) of the start-to-end frequency ratio. Error bars are 95%-confidence limits.

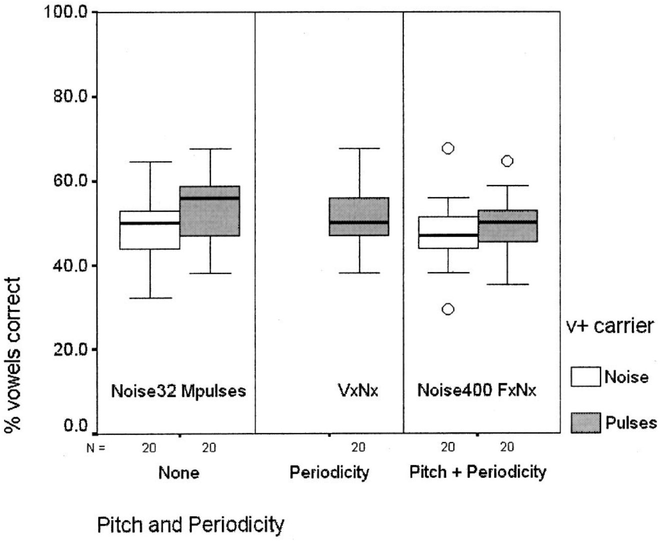

FIG. 5. Box and whisker plots showing percentage-correct vowel identification in the processor conditions. The legend ''v+ carrier'' indicates the carrier signal for voiced speech. Boxes represent the 25 to 75 percentile ranges of the data (over subjects, talker, and test run). In this and subsequent box and whisker plots, the bar within each box is the median. The whiskers show the range of scores excluding any outlying points that are more than 1.5 box widths from the box edge. Outliers are shown as open circles, or as asterisks for points more than 3 box widths from the box edge.

(ANOVA) was carried out on data from the last four test sessions, using factors of processor, talker, and test session. The only significant effect was that of test session [$F(3,12) = 14.34$, $p<0.001$, power=0.998].[4] Although all processors delivered equivalent representations of the slowly changing spectral structure of vowels, processors that signaled voice fundamental frequency and hence speaker sex might have been expected to show higher performance. This was not, however, the case. Performance with these four-channel processors was comparable to that found for a processor similar to *Noise400* in a previous study using the same vowel set from the female talker only (Rosen *et al.*, 1999).

C. Intervocalic consonants

1. Overall accuracy

Group results are shown in Fig. 6. A repeated-measures ANOVA of overall accuracy was carried out using factors of processor, talker, and test session. Here, there was no effect of test session, nor were there any significant interactions between any factors. A significant effect of talker [$F(1,4) = 43.9$, $p=0.003$, $\eta^2=0.916$, power=0.997][5] indicated higher scores for the female speech used here. There was a significant effect of processor [$F(4,16)=5.66$, $p=0.005$, power=0.926]. *A priori* contrasts comparing each processor to *Noise32* showed significantly higher scores for processor *Noise400* than for this reference condition ($p=0.025$, $\eta^2 = 0.754$, power=0.746). Hence, the use of a 400-Hz rather than a 32-Hz envelope bandwidth to modulate purely noise carriers increased performance. No other processor differed significantly from the reference, nor were other pairwise differences significant according to Bonferroni-corrected *post hoc* tests.

2. Consonant feature information

A second series of ANOVAs was performed on information transfer measures (Miller and Nicely, 1955) computed from confusion matrices summed over the last four test sessions. A summary of these ANOVAs is presented in Table II. The data are displayed in Fig. 7.

A more salient representation of periodic and aperiodic excitation would be expected to lead to improved identification of manner and voicing features (Faulkner *et al.*, 1989; Faulkner and Rosen, 1999). *A priori* comparisons of voicing information transmission with the reference condition *Noise32* showed significant differences for all four of the

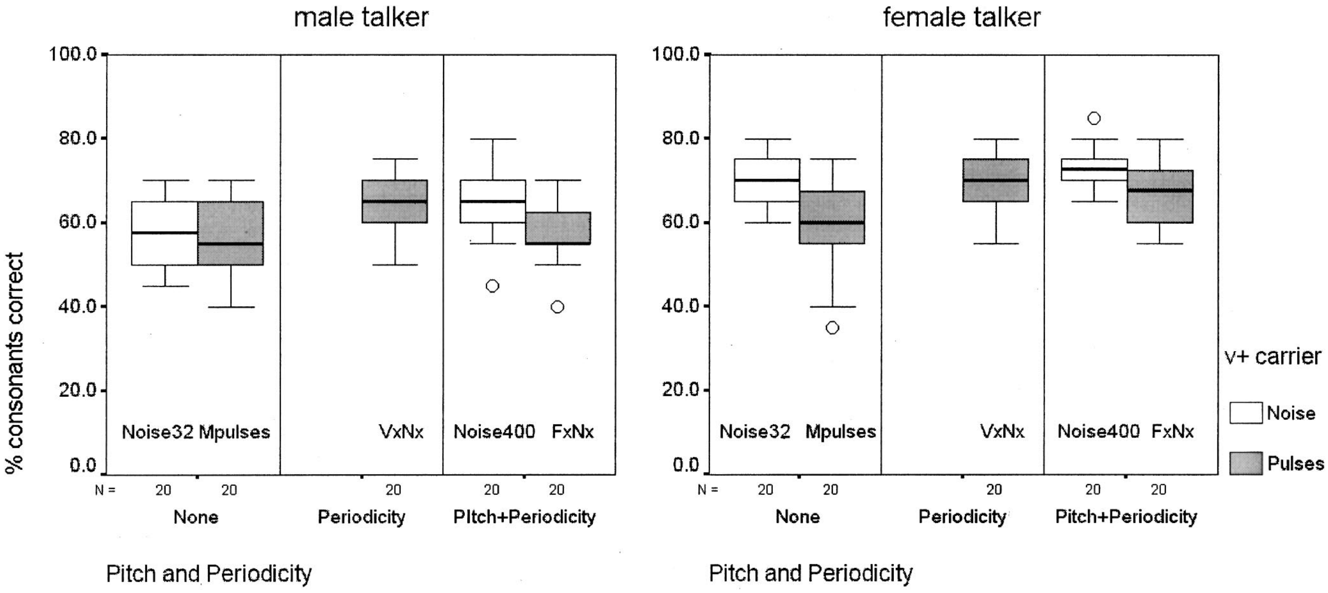

FIG. 6. Box and whisker plot showing percentage-correct consonant identification for each talker using the five processors. The box plots show the distribution of scores over subject and test run.

TABLE II. Summary of repeated-measures ANOVAs of consonant feature information transmission (interaction terms were always nonsignificant and are not shown).

Measure	Factor	*df*	*F*	*p*	η^2	Observed power
Voicing	Processor	1.22,4.88	19.10	0.007	0.827	0.948
	Talker	1,4	9.44	0.037	0.702	0.639
Place	Processor	4,16	8.27	0.001	0.674	0.988
	Talker	1,4	28.16	0.006	0.876	0.971
Manner	Processor	4,16	2.61	0.075	0.395	0.594
	Talker	1,4	11.96	0.026	0.749	0.735

other processors (see Table III). Scores were higher than the *Noise32* reference for processors *FxNx* and *VxNx* (both signaling periodicity information though the periodicity of the carrier), and for *Noise400*, (signaling periodicity information through higher rate envelope components). For processor *Mpulses*, voicing information scores were significantly lower than with the reference. Hence, all processors that represented the presence of speech periodicity, either by an explicit coding of periodicity and aperiodicity, or through the transmission of envelope modulations in the voice periodicity range, showed higher voicing transmission than the reference. The degree of voicing information provided by the

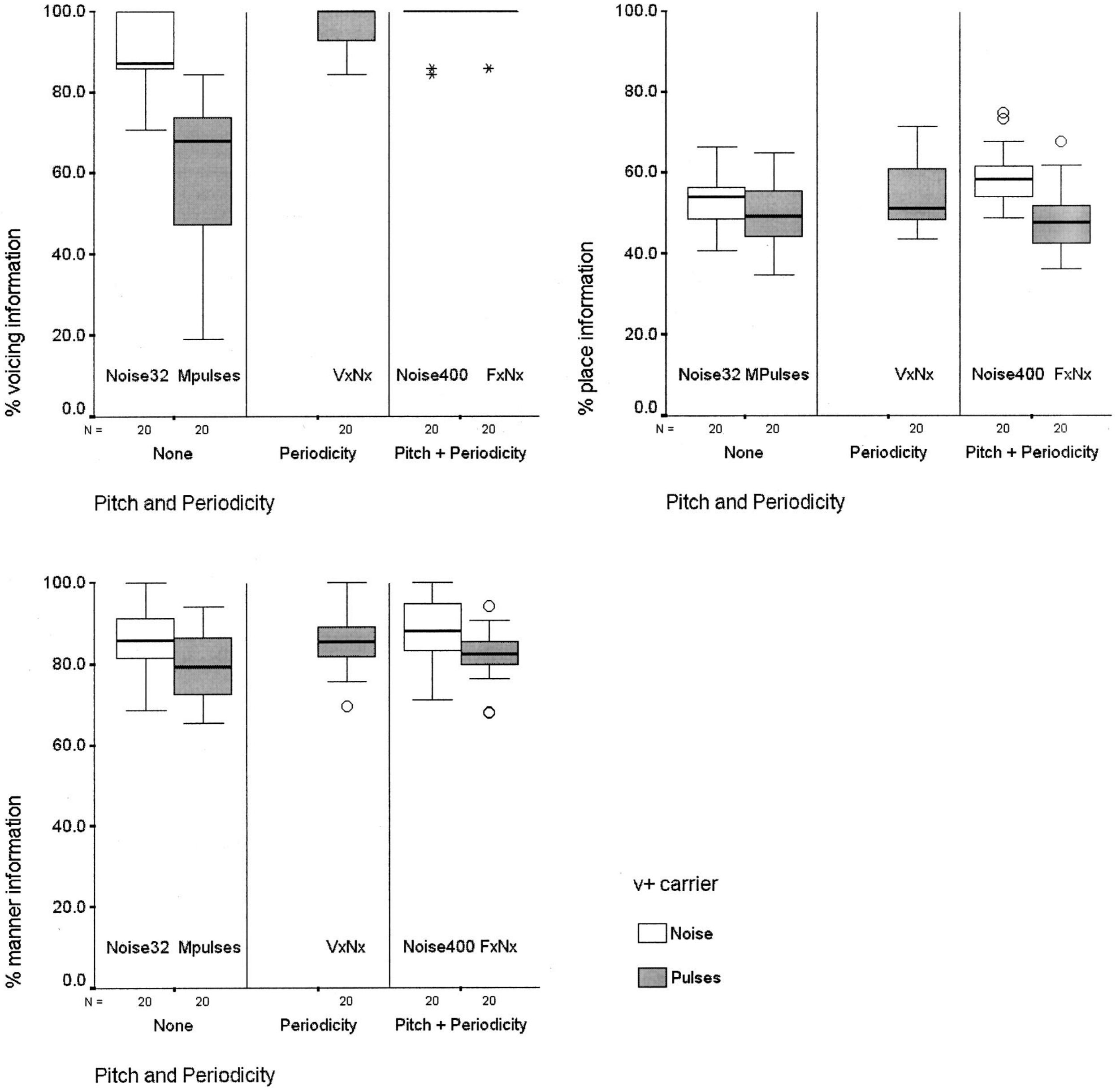

FIG. 7. Box and whisker plots of voicing, manner, and place information. The displayed data show the distribution of scores for each subject and talker.

TABLE III. Significant *a priori* contrasts against reference condition for consonant feature information. A + sign in the second column indicates scores higher than the reference, while − indicates lower scores.

Measure	Condition compared to *Noise32*	p	η^2	Observed power
Voicing	$FxNx+$	0.005	0.883	0.978
	Mpulses−	0.019	0.785	0.811
	Noise400+	0.027	0.745	0.727
	$VxNx+$	0.003	0.918	0.997
Place	$FxNx-$	0.008	0.854	0.944
	Noise400+	0.030	0.731	0.696

Noise32 reference is presumably based on dynamic spectral shape information. Processor *Mpulses* showed lower voicing information transmission than the *Noise32* reference while delivering identical dynamic spectral shape information carried by a constant and fixed-rate periodic carrier rather than by a noise carrier. This reduction of voicing information suggests that a carrier that is always periodic interferes with the use of spectral cues to this feature contrast.

For manner information there were no significant effects of processor, only an effect of talker, with higher scores for the female talker. This suggests that periodicity/aperiodicity is not a powerful cue for manner contrasts such as that between voiceless fricatives and voiced plosives or nasals, despite the difference in the excitation sources.

There were significant main effects of processor and talker for place information. An *a priori* comparison of processors against the *Noise32* reference showed two significant differences (Table III). Processor *Noise400* led to higher place information than the reference, while processor $FxNx$ gave significantly lower scores. A Bonferroni-corrected paired comparison between processors showed only one significant pairwise difference in place information scores, this being between processors *Noise400* and $FxNx$. All processors except for *Noise400* presented equivalent spectrotemporal information, while *Noise400* represented more rapid spectral envelope changes (resulting from the presence of envelope information above 32 and below 400 Hz) that were not present in the output of the other processors. This seems the most likely explanation for the higher place scores obtained through this processor. It is difficult to interpret the finding that place information with processor $FxNx$ was lower by 7% than with the *Noise32* reference.

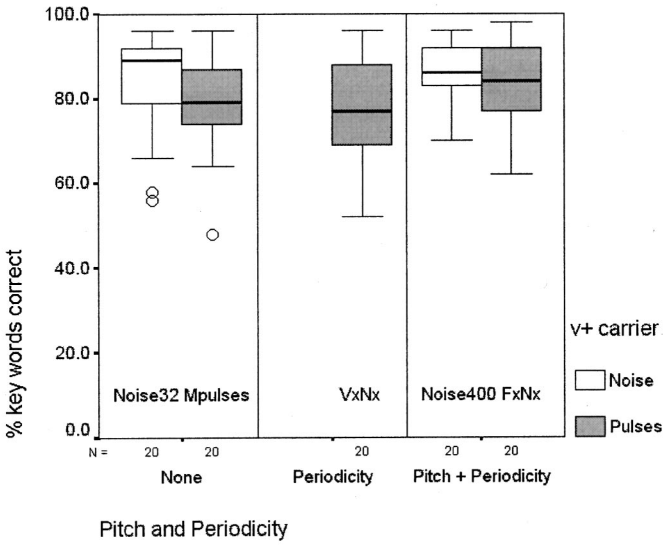

FIG. 8. Box and whisker plots of percentage of key words correctly identified from BKB sentences. Displayed data are the distribution of scores over subject and test session.

D. BKB sentences

Group scores using the key-word loose scoring method are shown in Fig. 8. Scores were similar in all conditions. Scores were rather high for a four-channel processor compared to another study that used the same materials and a processor similar to the *Noise400* condition (Rosen, Faulkner, and Wilkinson, 1999), and may be limited by ceiling effects. A repeated-measures ANOVA using factors of processor and test session was performed. There was no significant effect of test session. There was a significant main effect of processor [$F(0.128,0.009)=5.449$, $p=0.014$, power$=0.825$]. *A priori* contrasts showed no significant differences compared to the *Noise32* reference. Bonferroni-corrected paired comparisons between all five processors showed only one pairwise difference, this being between the highest-scoring processor *Noise400* and the lowest, $VxNx$.

E. Connected discourse tracking

Tracking rates through the four processors used for CDT (see Fig. 9) were all significantly lower than that with unprocessed speech (the $VxNx$ processor was not used here). A repeated-measures ANOVA was applied to CDT rates over the last four 10-min testing blocks with each processor, excluding the unprocessed speech condition. This showed a significant effect of block [$F(1.96,9.78)=8.22$, $p=0.008$, power$=0.875$]. Block did not interact with any other factor. An *a priori* contrast with the reference processor *Noise32* showed that rates through processor *Noise400* were significantly higher than rates obtained from the reference ($p=0.003$, $\eta^2=0.845$, power$=0.983$). Only this pairwise difference between processors was significant in *posthoc* Bonferroni comparisons.

That the noise-carrier processors showed a significant effect of the envelope filter cutoff suggests that speech-derived pitch and periodicity cues may increase the ease and rate of speech communication. However, this explanation would also require that rates through processor $FxNx$ (where the carrier conveys voice fundamental frequency) should exceed those through the fixed-pulse rate processor *Mpulses*. This, however, was not the case. It is concluded, therefore, that the difference between CDT rates through the *Noise400* and *Noise32* processors is due to the signaling of more rapid spectral changes by processor *Noise400* rather than to the presence of pitch and periodicity cues.

VI. DISCUSSION AND CONCLUSIONS

A. Salience of pitch and periodicity information across processors

Results from the frequency-glide labeling task confirm that processors differed in the salience of pitch information.

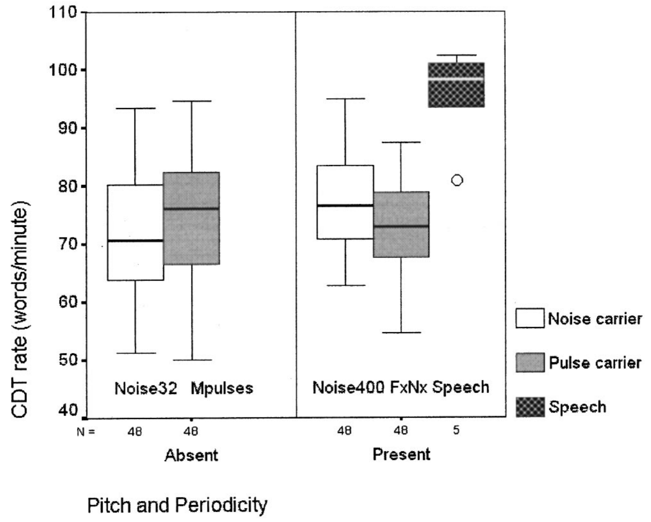

FIG. 9. Box and whisker plots of CDT rates (over subject and test session) for four processors and for unprocessed speech. The carrier used for voiced speech is indicated by box color. Data for speech are missing for one subject.

It can be assumed that temporal cues to speech periodicity/aperiodicity will have the same relative salience as those to voice pitch information across processors, since periodicity information necessarily resides in the same modulation frequency range as voice pitch.

The noise-carrier processors permitted relatively poor discrimination of pitch glide direction compared to a pulse train, as would be expected from previous studies of pitch percepts from amplitude-modulated noise. Even when the noise was modulated by an envelope having a 32-Hz envelope, discrimination of pitch glide direction was above chance performance. This is attributed to spectral envelope shifts that arise as harmonics of the input to the processor shift from one analysis band to the next. It seems unlikely, however, that with a signal such as speech, whose spectrum is constantly changing, there exist spectral shifts that are sufficiently well correlated with fundamental frequency to signal voice pitch change in the absence of more salient temporal cues or of resolved harmonic components.

Processors *Noise32*, and *Mpulses* differed only in the use of a noise compared to a pulsatile carrier, and apart from the random nature of the noise carrier, they conveyed identical spectral and temporal information, with temporal cues to pitch variation in the input signal being negligible. However, scores from processor *Mpulses* were significantly lower than from processor *Noise32* for the frequency-glide task. This suggests that a carrier with a strongly salient and constant pitch may in some way ''mask'' input-related pitch cues carried in the spectral information provided by the processor.

B. The role of pitch variation signaled by simulated processors

The processor with an *Fx*-controlled pulse rate never led to significantly higher speech scores than the noise-carrier processors carrying modulation up to 400 Hz. Despite the limited salience of informative pitch variation, we therefore conclude that 400 Hz envelope modulated noise carriers are adequate in this respect for the simulation of cochlear implant processors for speech intelligibility tasks such as those used here. Conversely, the limited sensitivity of any of the speech measures here to variations of pitch salience may signal the inadequacy of all of these measures for evaluating the availability to a listener of the full range of significant acoustic factors in speech perception.

Processors *FxNx* and *VxNx*, varying purely in the representation of informative pitch information, showed no significant differences in vowel, consonant, or sentence identification, indicating that with these tasks, pitch variation has no significance in the presence of spectral information. These two processors were not compared in CDT. However, there were no significant differences in CDT rate between the *FxNx* and *Mpulses* processors, indicating that neither pitch variation nor the periodic/aperiodic contrast contributed substantially in this task.

C. Consonant feature information from periodic and aperiodic carriers

The representation of periodicity and aperiodicity in the carrier signals does have a measurable effect on the transmission of consonant voicing. Compared to processor *Noise32*, consonant voicing information was significantly higher for processors that signaled the presence of periodicity in speech, whether through a change in the carrier's periodicity (*FxNx* and *VxNx*) or through noise-carrier modulations in the *Fx* range (*Noise400*). This outcome is consistent with a recently reported trend towards higher voicing information transmission in cochlear implant users as envelope bandwidths were increased from 40 to 320 and to 640 Hz (Fu and Shannon, 2000). Fu and Shannon also reported no effect of increasing the envelope bandwidth from 20 up to 640 Hz on overall or feature level consonant identification in simulations with normal listeners. However, the present data do show small but significant increases in both voicing information (8.4%) and overall consonant identification (4.7%) for a 400-Hz envelope bandwidth compared to one of 32 Hz. While the higher voicing transmission from processor *Noise400* could be due to relatively rapid between-channel level changes, the increased voicing information from processors *FxNx* and *VxNx* compared to *Noise32* can only be due to the encoding of periodicity, since these three processors all have the same 32-Hz envelope bandwidth.

Voicing information from processor *Mpulses*, where the carrier was always periodic, was significantly lower than from all the other processors. Since processor *Mpulses* differed from processor *Noise32* only in the use of a fixed-rate pulse carrier rather than noise, lower voicing scores from processor *Mpulses* must be attributable to the strong and constant periodic percept of the carrier, this being unrelated to the periodicity of the input. It appears that listeners do not readily associate this percept with voiceless speech. In contrast, it seems that the constant aperiodic percept from processor *Noise32* can be interpreted as representing voiced speech.[6] That this is possible may perhaps be based on our natural experience of whispered speech.

D. Results in relation to signals lacking spectral information

In the absence of spectral information, previous studies have shown that voice pitch information contributes substantially to the audio-visual perception of sentences and CDT (e.g., Rosen *et al.*, 1981; Waldstein and Boothroyd, 1994). When, as here, there is a limited degree of spectral information present, neither sentence perception nor CDT show clear evidence of a contribution of pitch information, despite the previous findings of strong effects of pitch information when spectral cues are absent.[7]

The auditory identification of consonants from spectrally invariant auditory signals also shows a substantial contribution from input-related periodicity or aperiodicity to contrasts of manner and voicing (Faulkner and Rosen, 1999). For voicing contrasts, the same effect of periodicity is evident here. For manner contrasts, however, there is no measurable contribution of periodicity information. It seems, then, that when limited spectral structure is present, spectral balance cues are sufficient to mark those manner of articulation differences that can also be signaled by temporal cues to speech periodicity/aperiodicity.

E. The role of pitch in speech communication

The present studies are likely to substantially underestimate the contribution of pitch information to communication, especially where paralinguistic cues (e.g., to talker identity or pragmatics) are important. Furthermore, envelope-based pitch cues have been shown to contribute to Chinese sentence perception through similar processors (Fu *et al.*, 1998). The most reasonable interpretation of our findings is not that factors such as voice pitch are unimportant. Rather, we would argue that the intelligibility measures used here lack sensitivity to important aspects of speech quality. Since the speech tests used here are, with the exception of CDT, essentially the same as those almost universally used in clinical research evaluating cochlear implant benefit, it may be that conventional speech-based measures of benefit are missing aspects of speech perception that are of real importance in speech communication. Intonation is widely held to be a major factor in the development of spoken language. Hence, the role of voice pitch information in cochlear implant speech processing should not be dismissed simply because it appears to have little impact on intelligibility for adult listeners. It is entirely possible that during speech development, intonation and other prosodic factors may play a much larger role in perceptual speech processing than in the mature adult.

Finally, we note that these simulation data suggest that the use of a fixed-rate carrier signals in the voice Fx range (here at 150 Hz) as carriers of multiband speech envelope information may be inappropriate in speech perceptual prostheses because of the inherent periodicity and fixed-pitch percepts produced by such carriers. Compared to aperiodic carriers, or carriers signaling speech-derived periodicity and aperiodicity, the identification of consonant voicing contrasts is significantly poorer. This difficulty may not arise with the higher pulse rates that are typically used in CIS processors, but it does seem likely to limit the effectiveness of cochlear implant speech processors that use fixed-pulse rates within the voice fundamental frequency range.

ACKNOWLEDGMENTS

Supported by a Wellcome Trust Summer Vacation Scholarship to Clare Smith (Ref. No. VS/98/UCL/001) and CEC TIDE project OSCAR (No. TP 1217). We are grateful to Chris Turner, Philip Loizou, and an anonymous reviewer for helpful comments on a previous version of this paper.

[1] The modulation of sinusoidal carriers by envelopes whose bandwidths extend into the voice fundamental frequency range leads to a rather complex acoustic stimulus, due to the presence of sidebands. This results in spectral cues to pitch even though the spectra are not harmonic.

[2] Data described by Dorman *et al.* (1996) do indicate weak spectrally based pitch percepts for one user of a CIS processor when the input signals were single sine waves. For sinusoidal stimuli, the spectral envelope as represented by the processor filter bank is correlated to input fundamental frequency.

[3] The amplitude modulation of pulsatile carriers inevitably affects spectral detail through the introduction of sidebands. Since the modulating bandwidth for such carriers was limited to 32 Hz, the spectrum at each harmonic component will be only slightly broadened. Such details are not expected to be perceptually significant.

[4] Here and elsewhere, F tests on factors with $df>1$ are based on Huynh–Feldt epsilon correction factors.

[5] η^2 indicates the eta-squared statistic that represents the proportion of variability in the dependent variable due to the independent variable.

[6] A related suggestion, that a fixed rate of pulsatile electrical stimulation within the voice fundamental frequency range may ''interfere'' with envelope perception, has recently been made (Fu and Shannon, 2000).

[7] It remains possible, although rather implausible, that a contribution of pitch variation to CDT performance occurs only with audio-visual presentation.

Breeuwer, M., and Plomp, R. (**1986**). ''Speech reading supplemented with auditorily presented speech parameters,'' J. Acoust. Soc. Am. **79**, 481–499.

DeFilippo, C. L., and Scott, B. L. (**1978**). ''A method for training and evaluation of the reception of ongoing speech,'' J. Acoust. Soc. Am. **63**, 1186–1192.

Dorman, M. F., Smith, L. M., Smith, M., and Parkin, J. L. (**1996**). ''Frequency discrimination and speech recognition by patients who use the Ineraid and continuous interleaved sampling cochlear-implant signal processors,'' J. Acoust. Soc. Am. **99**, 1174–1184.

Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997a**). ''Simulating the effect of cochlear-implant electrode insertion depth on speech understanding,'' J. Acoust. Soc. Am. **102**, 2993–2996.

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

Dudley, H. (**1939**). ''The vocoder,'' Bell Lab. Rec. **17**, 122–126.

EPI Group (**1986**). *The BKB (Bamford-Kowal-Bench) Standard Sentence Lists* [Video recordings] (Department of Phonetics and Linguistics, University College London, London).

Faulkner, A., Potter, C., Ball, G., and Rosen, S. (**1989**). ''Audiovisual speech perception of intervocalic consonants with auditory voicing and voiced/voiceless speech pattern presentation,'' Speech, Hearing and Language, Work in progress, University College London, Department of Phonetics and Linguistics **3**, 85–106.

Faulkner, A., and Rosen, S. (**1999**). ''Contributions of temporal encodings of voicing, voicelessness, fundamental frequency and amplitude variation in audio-visual and auditory speech perception,'' J. Acoust. Soc. Am. **106**, 2063–2073.

Foster, J. R., Summerfield, A. Q., Marshall, D. H., Palmer, L., Ball, V., and

Rosen, S. (**1993**). ''Lip-reading the BKB sentence lists; corrections for list and practice effects,'' Br. J. Audiol. **27**, 233–246.

Fourcin, A. J., Rosen, S. M., Moore, B. C. J., Douek, E. E., Clarke, G. P., Dodson, H., and Bannister, L. H. (**1979**). ''External electrical stimulation of the cochlea: Clinical, psychophysical, speech-perceptual and histological findings,'' Br. J. Audiol. **13**, 85–107.

Fu, Q.-J., and Shannon, R. V. (**2000**). ''Effect of stimulation rate on phoneme recognition by Nucleus-22 cochlear implant listeners,'' J. Acoust. Soc. Am. **107**, 589–597.

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

Grant, K. W., Ardell, L. H., Kuhl, P. K., and Sparks, D. W. (**1985**). ''The contribution of fundamental frequency, amplitude envelope and voicing duration cues to speechreading in normal-hearing subjects,'' J. Acoust. Soc. Am. **77**, 671–677.

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

McDermott, H. J., and McKay, C. M. (**1997**). ''Musical pitch perception with electrical stimulation of the cochlea,'' J. Acoust. Soc. Am. **101**, 1622–1631.

Miller, G. A., and Nicely, P. E. (**1955**). ''An analysis of perceptual confusions among some English consonants,'' J. Acoust. Soc. Am. **27**, 338–352.

Pollack, I. (**1969**). ''Periodicity pitch for white noise—fact or artefact,'' J. Acoust. Soc. Am. **45**, 237–238.

Risberg, A. (**1974**). ''The importance of prosodic speech elements for the lipreader,'' Paper presented at the Visual and Audio-Visual Perception of Speech, Sixth Danavox Symposium: Scandinavian Audiology, Supplementum 4.

Risberg, A., and Lubker, J. L. (**1978**). ''Prosody and speechreading,'' Report of STL-QPSR, Dept. of Linguistics, University of Stockholm, Stockholm, Sweden, **4**, pp. 1–16.

Rosen, S., Faulkner, A., and Wilkinson, L. (**1999**). ''Perceptual adaptation by normal listeners to upward shifts of spectral information in speech and its relevance for users of cochlear implants,'' J. Acoust. Soc. Am. **106**, 3629–3636.

Rosen, S., Fourcin, A. J., and Moore, B. C. J. (**1981**). ''Voice pitch as an aid to lipreading,'' Nature (London) **291**, 150–152.

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

Shannon, R. V., Zeng, F.-G., and Wygonski, J. (**1998**). ''Speech recognition with altered spectral distribution of envelope cues,'' J. Acoust. Soc. Am. **104**, 2467–2476.

Waldstein, R. S., and Boothroyd, A. (**1994**). ''Speechreading enhancement using a sinusoidal substitute for voice fundamental frequency,'' Speech Commun. **14**, 303–312.

Wilson, B., Finley, C., Lawson, D., Wolford, R., Eddington, D., and Rabinowitz, W. (**1991**). ''Better speech recognition with cochlear implants,'' Nature (London) **352**, 2.

Wilson, B., Zerbi, M., Finley, C., Lawson, D., and van den Honert, C. (**1997**). Eighth Quarterly Progress Report, 1 May through 31 July 1997. NIH Project N01-DC-5-2103: Speech Processors for Auditory Prostheses: Research Triangle Institute.