

Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers

Michael K. Qin^{a)} and Andrew J. Oxenham^{b)}

Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 and Harvard-MIT Division of Health Sciences and Technology, Speech and Hearing Bioscience and Technology Program, Cambridge, Massachusetts 02139

(Received 8 November 2002; revised 30 March 2003; accepted 14 April 2003)

This study investigated the effects of simulated cochlear-implant processing on speech reception in a variety of complex masking situations. Speech recognition was measured as a function of target-to-masker ratio, processing condition (4, 8, 24 channels, and unprocessed) and masker type (speech-shaped noise, amplitude-modulated speech-shaped noise, single male talker, and single female talker). The results showed that simulated implant processing was more detrimental to speech reception in fluctuating interference than in steady-state noise. Performance in the 24-channel processing condition was substantially poorer than in the unprocessed condition, despite the comparable representation of the spectral envelope. The detrimental effects of simulated implant processing in fluctuating maskers, even with large numbers of channels, may be due to the reduction in the pitch cues used in sound source segregation, which are normally carried by the peripherally resolved low-frequency harmonics and the temporal fine structure. The results suggest that using steady-state noise to test speech intelligibility may underestimate the difficulties experienced by cochlear-implant users in fluctuating acoustic backgrounds. © 2003 Acoustical Society of America. [DOI: 10.1121/1.1579009]

PACS numbers: 43.66.Ts, 43.71.Ky, 43.66.Dc [PA]

I. INTRODUCTION

Speech has been shown to be a very robust medium for communicating information (Fletcher and Galt, 1950; Miller and Licklider, 1950; Remez *et al.*, 1994; Stevens, 1998). Although the precise mechanisms underlying the apparent resilience to interference and distortion are still not well understood, the ability of speech to convey information under adverse conditions is generally attributed to the layers of acoustic, phonetic, and linguistic redundancies. Shannon *et al.* (1995), using a noise-excited vocoder,¹ provided a dramatic demonstration of these redundancies at work. They found that despite a severe reduction in spectral cues and the elimination of temporal fine-structure information, sentences presented in the absence of interfering sounds could be recognized with as few as four frequency bands. Subsequent studies have shown that while more frequency bands are needed for speech reception in steady-state noise, good sentence recognition is still possible at relatively low signal-to-noise ratios (e.g., Dorman *et al.*, 1998).

The processing schemes used in these studies are designed to simulate the effects of cochlear-implant stimulation (Wilson *et al.*, 1991). They can therefore be used to provide insights into the relative efficacy of different processing algorithms without using valuable implantee testing time (Blamey *et al.*, 1984). Indeed, at least for low numbers of frequency bands, acoustic simulations of cochlear-implant processing using normal-hearing listeners have yielded results that are reasonably comparable to those of actual implant patients (Friesen *et al.*, 2001; Carlyon *et al.*, 2002). An

other use for such schemes is to probe the acoustic features necessary for speech reception in normal-hearing listeners. A number of studies indicate that important information is carried in the envelopes of the stimulus after filtering into frequency sub-bands (Houtgast *et al.*, 1980; Drullman, 1995; Smith *et al.*, 2002). From the results obtained so far, it may be concluded that speech reception requires minimal frequency selectivity and no temporal fine-structure information. This conclusion seems at odds with the experiences of many hearing-impaired listeners.

While hearing-impaired listeners often perform well in quiet conditions (when audibility is corrected for with amplification), many experience great difficulty in noisy conditions. The difference in performance between normal-hearing and hearing-impaired listeners is especially pronounced in temporally fluctuating maskers and maskers with spectral notches (Festen and Plomp, 1990; Gustafsson and Arlinger, 1994; Peters *et al.*, 1998). In particular, while normal-hearing listeners show large improvements in speech reception when spectral and/or temporal fluctuations are introduced into a masker, hearing-impaired listeners often show much less benefit (Festen and Plomp, 1990; Peters *et al.*, 1998). It is thought that normal-hearing listeners are able to make use of the improved local target-to-masker ratio in the masker's spectral and temporal dips. In contrast, hearing-impaired listeners, with their poorer frequency selectivity (Patterson *et al.*, 1982; Glasberg and Moore, 1986) and poorer effective temporal resolution (Glasberg and Moore, 1992; Oxenham and Moore, 1997), may be less able to benefit from the improved local target-to-masker ratio found in the spectral and temporal dips of the masker.

^{a)}Electronic mail: qin@mit.edu

^{b)}Electronic mail: oxenham@mit.edu

In the case of cochlear implants and implant simulations, the finding that better frequency resolution (i.e., a greater number of frequency bands) is required for speech reception in noise than in quiet (Dorman *et al.*, 1998; Fu *et al.*, 1998) parallels the finding that spectral smearing is more detrimental to speech reception in noise than in quiet (ter Keurs *et al.*, 1992; Baer and Moore, 1993). It is also consistent with the hypothesized effect of poorer frequency selectivity in hearing-impaired listeners. The perceptual effect of eliminating the temporal fine structure in cochlear-implant simulations is less clear. Pitch perception and the ability to discriminate different fundamental frequencies (F_0 s), is thought to rely primarily on fine-structure information, in particular the information carried in peripherally resolved, lower-order harmonics (e.g., Plomp, 1967; Houtsma and Smurzynski, 1990; Smith *et al.*, 2002). While the envelopes of implant-processed stimuli carry some periodicity information, the pitch salience associated with such envelope periodicity is rather weak (Burns and Viemeister, 1976; 1981; Shackleton and Carlyon, 1994).

Fundamental frequency information has long been thought to play an important role in perceptually segregating simultaneous and nonsimultaneous sources (Brokx and Nooteboom, 1982; Assmann and Summerfield, 1990; 1994; Bird and Darwin, 1998; Vliegen and Oxenham, 1999; see Darwin and Carlyon, 1995 for a review). A reduction in F_0 cues produced by cochlear-implant processing may lead to greater difficulties in segregating different sources. If the perception of implant-processed speech is based on envelope fluctuations, as suggested above, then listeners must accurately distinguish the envelope fluctuations of the target from those of the masker. Similarly, a listener can only take advantage of spectral and temporal dips in the masker if the listener can accurately identify the presence of the dips.

The aim of the present study was to investigate the effects of fluctuating maskers on the reception of simulated implant-processed speech. We hypothesized that the reduction in F_0 cues produced by the implant simulations would particularly affect conditions where the ability to discriminate the target from the masker is thought to play an important role in determining speech reception thresholds (e.g., speech in the presence of competing talkers or fluctuating backgrounds). Speech reception was measured in normal-hearing listeners as a function of target-to-masker ratio, processing condition (4, 8, or 24 channels, or unprocessed) and masker type (steady-state speech-shaped noise, speech-shaped noise modulated with a speech envelope, single male talker, and single female talker).

II. METHODS

A. Subjects

Thirty-two normal-hearing listeners (15 females) with audiometric thresholds of less than 20 dB HL at octave frequencies between 125 and 8000 Hz, participated in this study. Their ages ranged from 18 to 46 (median age 22). They were all native speakers of American English.

B. Stimuli

All stimuli in this study were composed of a target sentence presented in the presence of a masker. The stimulus tokens were processed prior to each experiment. The targets and maskers were combined at the desired target-to-masker ratios (TMRs) prior to any processing. TMRs were computed based on the token-length root-mean-square (rms) amplitudes of the signals. Maskers were gated on and off with 250-ms raised-cosine ramps 250 ms prior to and 250 ms after the end of each target sentence.

The targets were H.I.N.T. sentences (Nilsson *et al.*, 1994) spoken by a male talker. The H.I.N.T sentence corpus consists of 260 phonetically balanced high-context sentences of easy-to-moderate difficulty. Each sentence is composed of four to seven keywords.

Since differences in the F_0 of voicing are thought to contribute to speaker segregation (Brokx and Nooteboom, 1982; Assmann and Summerfield, 1990; 1994; Darwin and Carlyon, 1995; Bird and Darwin, 1998), we chose a male single-talker masker with a mean F_0 (111.4 Hz) similar to that of the target talker (110.8 Hz) and a female single-talker masker with a mean F_0 (129.4 Hz) almost 3 semitones higher. The motivation for using different gender single-talker interferers came from the observation that normal-hearing listeners benefit from F_0 differences between target and interfering talkers (Brokx and Nooteboom, 1982; Assmann and Summerfield, 1990; 1994; Bird and Darwin, 1998). Talker F_0 s were estimated using the YIN program provided by de Cheveigné and Kawahara (2002). The male single-talker maskers were excerpts from the audio book *Timeline* (novel by M. Crichton) read by Stephen Lang. The female single-talker maskers were excerpts from the audio book *Violin* (novel by A. Rice) read by Maria Tucci. To avoid long silent intervals in the masking speech, such as sentence-level pauses, both single-talker maskers were automatically preprocessed to remove silent intervals greater than 100 ms. The maskers were then subdivided into nonoverlapping segments to be presented at each trial.

The single-talker maskers and speech-shaped-noise masker were spectrally shaped to match the long-term power spectrum of the H.I.N.T. sentences. The amplitude-modulated speech-shaped noise masker was generated by amplitude modulating the steady-state speech-shaped noise with the broadband speech envelope of the male single-talker masker (lowpass filtered at 50 Hz; first-order Butterworth filter).

For a given listener, the target sentence lists were chosen at random, without replacement, from among the 25 lists of H.I.N.T. sentences. This was done to ensure that no target sentence was presented more than once to any given listener. Data were collected using one list (i.e., ten sentences) for each TMR.

C. Stimulus processing

All stimulus tokens were processed prior to each experiment. The cochlear-implant simulator was implemented using MATLAB (Mathworks, Natick, MA) in the following manner. The stimuli (target plus masker) were first bandpass

TABLE I. Target-to-masker ratios (TMR) used in the study. The values in the table represent the minimum, maximum, and step size of the TMR (in dB). The step sizes are in parentheses.

Processing condition	Masker type	Target-to-masker ratio (dB)
Unprocessed	Male interference	-20 to 5 (5)
	Female interference	-20 to 5 (5)
	Modulated noise	-25 to 0 (5)
	Steady-state noise	-15 to 0 (3)
24 channels	Male interference	-15 to 10 (5)
	Female interference	-15 to 10 (5)
	Modulated noise	-20 to 5 (5)
	Steady-state noise	-10 to 10 (4)
8 channels	Male interference	-5 to 20 (5)
	Female interference	-5 to 20 (5)
	Modulated noise	-10 to 15 (5)
	Steady-state noise	-5 to 20 (5)
4 channels	Male interference	5 to 30 (5)
	Female interference	5 to 30 (5)
	Modulated noise	5 to 30 (5)
	Steady-state noise	5 to 30 (5)

filtered (sixth-order Butterworth filters) into 4, 8, or 24 contiguous frequency bands (or channels) between 80 and 6000 Hz. The entire frequency range was divided equally in terms of the Cam scale² (Glasberg and Moore, 1990). The 3-dB channel bandwidths were approximately 6.98 Cams, 3.49 Cams, and 1.16 Cams for the 4-, 8-, and 24-channel conditions, respectively. The envelopes of the signals were extracted by half-wave rectification and lowpass filtering (using a second-order Butterworth filter) at 300 Hz, or half the bandpass filter bandwidth, whichever was lower. The 300-Hz cutoff frequency was chosen to preserve, as far as possible, F_0 cues in the envelope. The envelopes were then used to modulate narrowband noises, filtered by the same bandpass filters that were used to filter the original stimuli. Finally, the modulated narrowband noises were summed and scaled to have the same level as the original stimuli.

D. Procedure

The 32 listeners were divided into four groups of eight. Each group was tested on only one of the four processing conditions (i.e., 4, 8, 24 channels, or unprocessed). The speech reception of each listener was measured in the presence of all four masker types (single male and female talkers, modulated and steady-state speech-shaped noise), at six TMRs (see Table I). The TMRs for each processing condition and masker type were determined in an earlier pilot study, using two to three listeners. The TMRs for each experimental condition were set to avoid floor and ceiling effects in the psychometric function.

The target and masker were combined at the appropriate TMR, processed, and stored on disk prior to the experiments. The processed stimuli were converted to the analog domain using a soundcard (LynxStudio, LynxOne) at 16-bit resolution with a sampling rate of 22 050 Hz. The stimuli were then passed through a headphone buffer (TDT HB6) and presented diotically at 60 dB SPL via Sennheiser HD580 headphones to the listener seated in a double-walled sound-

insulation booth. Listeners typed their responses into a computer via the keyboard.

For practice, the listeners were presented with 20 stimuli, five from each of the four masking conditions. In each practice masking condition, the target sentences were presented at four different TMRs. The target sentences used in the practice session were from the Harvard-Sentence database (IEEE, 1969). The practice sessions were designed to acclimate the listeners to the processed stimuli. Feedback was given during the practice sessions, but not during the experimental sessions.

E. Analysis

Listener responses were scored offline by the experimenter. Each listener's responses for a given TMR, under a given masker condition, were grouped together to produce a percent-correct score. Keywords were used to calculate the percent correct. Obvious misspellings of the correct word were considered correct.

III. RESULTS

A. Fits to the psychometric functions

The percent-correct scores as a function of TMR under a given masker condition for each listener were fitted to a two-parameter sigmoid model (a cumulative Gaussian function)

$$\text{Percent correct} = \frac{100}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\text{TMR}} \exp\left(-\frac{-(\text{TMR}-\text{SRT})^2}{2\sigma^2}\right), \quad (1)$$

where SRT is the speech-reception threshold³ (dB), σ is related to the slope of the function, and TMR is the target-to-masker ratio (dB). Figure 1 shows sample data from one listener, along with the best-fitting curve (heavy) according to Eq. (1). The other, lighter curves in the figure are the fits to the data from the other seven listeners in that experimental condition. The two-parameter model assumes that listeners' peak reception performance is 100%. This assumption may be valid for the 24- and 8-channel conditions, but it is probably not valid for the 4-channel condition. Therefore, the initial model had a third parameter, associated with the peak performance. However, the goodness of fit and the estimated SRTs of the three-parameter model were very similar to those of the two-parameter model, leading us to select the model with fewer parameters.

The two-parameter model provided generally good fits to the curves of performance as a function of TMR. Presented in Table II are the mean values of SRT, σ , and standard error of fit, averaged across listeners. The standard deviations are shown in parentheses. The individual standard errors of fit⁴ had a mean of 7.25% with a standard deviation of 3.7% (median of 7.01% and a worst case of 20.93%). Combined according to experimental conditions, the average standard errors of fit (Table II) were generally less than 10%.

B. Speech-reception thresholds

In general, performance across listeners was reasonably consistent, so only the mean SRT values as a function of masker condition and processing condition are plotted in Fig.

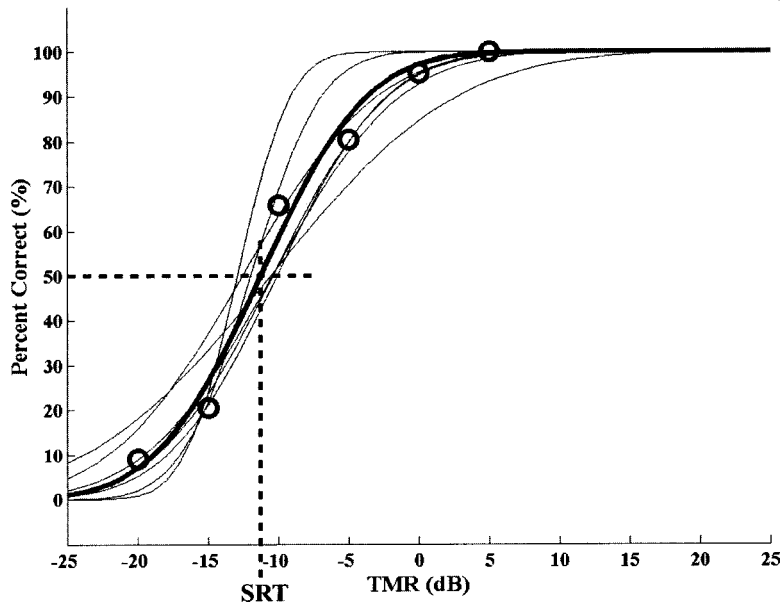


FIG. 1. An example of the two-parameter sigmoid model fitting procedure. The two-parameter sigmoid model (heavy line) is fitted to the speech reception performance data of an individual listener (open circles). The light lines are the functions fitted to the data of the other listeners in the same experimental condition (data not shown). The speech reception threshold (SRT) is the target-to-masker ratio (TMR), where 50% of the words were correctly identified.

2. The mean SRT values and standard errors of means (Table II) were derived from the SRT values of individual model fits. Since an SRT value is the TMR where 50% of the keywords are correctly identified, a higher SRT value implies a condition more detrimental to speech reception. Figure 2 shows that SRT values in all masker conditions were strongly affected by implant processing. As the number of spectral channels decreased, the SRT values under all masker types increased. However, the rate of increase differed between masker types. As the number of spectral channels decreased, SRT values increased faster in the presence of fluctuating maskers than in the presence of steady-state noise, particularly for the single-talker maskers.

A two-way mixed-design analysis of variance (ANOVA) was performed using STATISTICA (StatSoft, Tulsa, OK) to de-

termine the statistical significance of the findings, with SRT as the dependent variable, processing condition as the between-subjects factor, and masker type as the within-subjects factor. The ANOVA indicated that both main factors and their interaction were statistically significant (processing condition: $F_{3,28}=218.1$; masker type: $F_{3,84}=7.5$; interaction: $F_{9,84}=8.7$; $p<0.001$ in all cases). A *post hoc* test according to Fisher's LSD ($\alpha=0.05$) indicated several significant differences between the different experimental conditions, as outlined below.

In the unprocessed conditions, the steady-state noise masker was significantly more effective than any of the modulated maskers. However, under implant processing the reverse was true, with the exception of the 24-channel processed modulated speech-shaped noise condition. These differential effects are illustrated in Fig. 3, which treats the steady-state masker as the baseline condition and plots the differences in SRT values between the steady-state noise and the other maskers as a function of processing condition. Significant differences between SRTs in steady-state noise and those in the other conditions are labeled with asterisks in Fig. 3.

The single-talker interferers produced significantly higher SRT values than steady-state noise in all processed conditions (i.e., 24, 8, and 4 channels). In contrast, the modulated noise produced lower SRT values than the steady-state noise in the 24-channel condition, and was not significantly different from the steady-state noise in the 8- and 4-channel conditions. As illustrated in Fig. 2, in all conditions the transition from unprocessed to 24-channel processing resulted in a large increase in SRT value. This is despite the fact that the 24-channel condition represents frequency resolution approaching that found in normal-hearing listeners. This finding is explored further in Sec. IV C.

There was no significant difference in the SRT values between the male and female single-talker maskers in any processing condition (Fig. 2). Given our hypothesized effect

TABLE II. Mean sigmoidal model parameter values [Eq. (1)], averaged across listeners. The standard deviations are in parentheses. The standard error of fit provides a numerical indicator for how well the model fits the data. SRT is the speech reception threshold, and σ is related to the slope of the function.

Processing condition	Masker type	SRT (dB TMR)	σ	Standard error of fit (%)
Unprocessed	Male interference	-10.3(2.4)	7.5(1.9)	8.3(3.5)
	Female interference	-11.3(1.1)	6.3(2.3)	6.8(3.2)
	Modulated noise	-9.1(0.6)	4.1(1.5)	5.6(2.3)
	Steady-state noise	-6.7(0.8)	3.4(0.7)	5.5(3.4)
24 channels	Male interference	0.7(1.8)	5.1(1.1)	6.4(3.4)
	Female interference	0.6(1.8)	5.0(1.0)	4.8(2.1)
	Modulated noise	-3.3(0.6)	5.0(1.1)	4.0(2.3)
	Steady-state noise	-1.2(1.2)	3.3(0.6)	3.8(1.8)
8 channels	Male interference	6.4(2.2)	6.0(2.0)	9.1(2.0)
	Female interference	6.7(1.5)	5.2(1.4)	7.2(1.7)
	Modulated noise	4.6(2.1)	7.7(1.5)	6.8(3.4)
	Steady-state noise	4.2(0.8)	4.1(1.2)	5.4(1.9)
4 channels	Male interference	18.1(3.1)	14.1(1.8)	9.9(3.1)
	Female interference	18.3(4.3)	15.3(3.3)	12.5(3.8)
	Modulated noise	15.6(4.4)	18.7(4.8)	9.3(2.2)
	Steady-state noise	14.9(5.4)	19.3(9.2)	10.6(3.4)

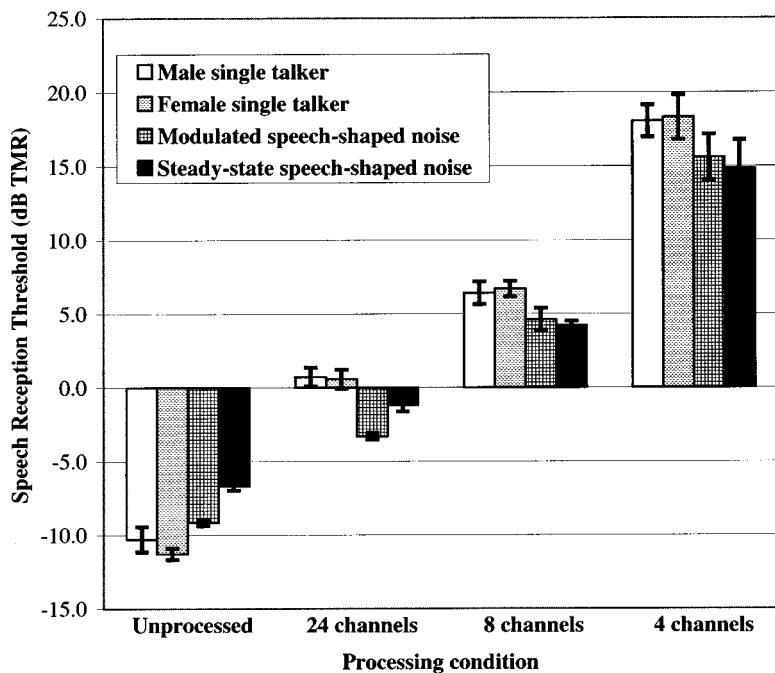


FIG. 2. Speech reception threshold, in terms of target-to-masker ratio, as a function of processing condition in the presence of a male single talker (unshaded), a female single talker (dotted), modulated noise (grid), and steady-state noise (solid). The plotted values and their respective standard deviations can be found in Table II.

of F_0 differences in source segregation, this may seem unexpected. This finding is explored further in Sec. IV D.

IV. DISCUSSION

A. Single-talker versus steady-state noise interference

Our results in the unprocessed conditions are consistent with previous studies in showing that SRT values are lower for single-talker interferers than for steady-state noise (e.g., Festen and Plomp, 1990; Peissig and Kollmeier, 1997; Peters *et al.*, 1998). The improved performance found with single-talker interference, relative to steady-state noise, has been ascribed to listeners' ability to gain information from temporal or spectral minima in the maskers. However, to make use of local masker minima, the listener must have cues to distinguish the target from masker. Voice F_0 is a generally ac-

cepted segregation cue for normal-hearing listeners (Brox and Nootboom, 1982; Bird and Darwin, 1998; Freyman *et al.*, 1999; Brungart, 2001). Our hypothesis was that the reduction in F_0 cues, produced by simulated cochlear-implant processing, would particularly affect speech reception where the ability to discriminate the target from the masker is thought to play an important role. The results from the processed conditions are consistent with the hypothesis: not only are the benefits of spectral and temporal masker dips eliminated, but the single-talker interferers go from being the least effective maskers in the unprocessed conditions to being the most effective maskers in all the processed conditions (see Figs. 2 and 3).

B. Modulated noise versus steady-state noise

If simulated cochlear-implant processing led to a global inability to use temporal minima in fluctuating maskers, the

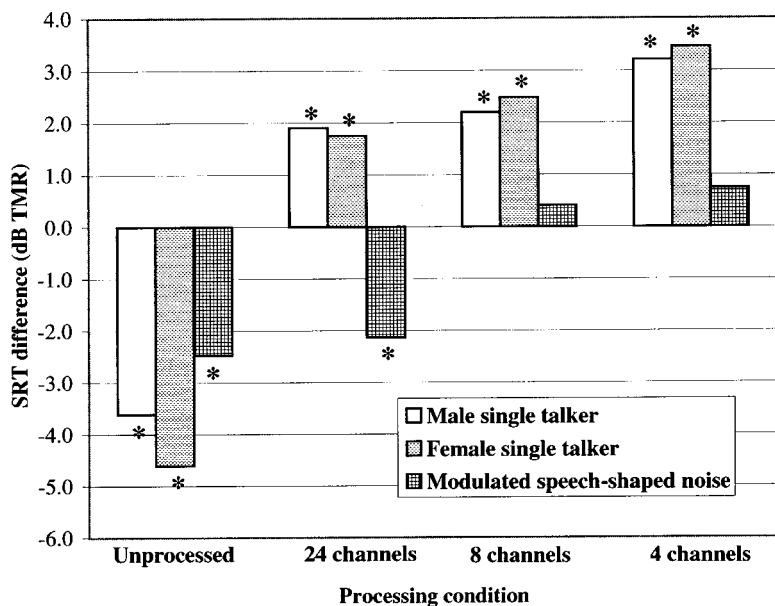


FIG. 3. SRT differences between the steady-state noise masker and the male single-talker (unshaded), female single-talker (dotted), and modulated-noise (grid) maskers are shown as a function of processing condition. Masked thresholds significantly different from those in the steady-state noise, according to Fisher's LSD test ($\alpha=0.05$), are labeled by an asterisk.

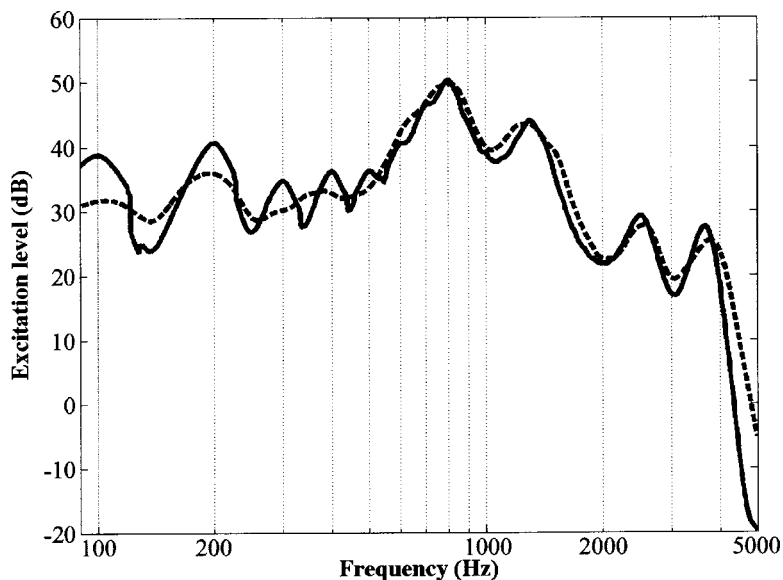


FIG. 4. An illustration of the difference in effective spectral resolution between unprocessed (solid) and 24-channel processed (dashed) conditions. This figure is the output of the excitation pattern model (Moore *et al.*, 1997) in response to a 500-ms Klatt synthesized vowel /a/ with an F_0 of 100 Hz (Klatt, 1980).

same deterioration in performance would be expected in the modulated-noise masking conditions as was found for the single-talker interferers. In fact, the difference in performance between the modulated-noise and the steady-state-noise conditions remains roughly constant for the unprocessed and 24-channel conditions. Even for 8 and 4 channels, SRT values in the modulated-noise conditions are not significantly higher than in the steady-state noise conditions.

Without F_0 cues, listeners may still maintain high levels of speech reception in the presence of interference by utilizing different cues. For example, when speech is presented in the presence of steady-state noise, a listener may be able to use common slowly varying envelope modulation as a cue for segregating the target from the noise, as most slow-varying envelope modulations will belong to the target. In the case of the amplitude-modulated speech-shaped noise masker, the noise is always modulated coherently across frequency. Speech, on the other hand, does not always modulate coherently across all frequencies. Listeners could use the more consistent comodulation of the amplitude-modulated noise as a cue for source segregation. However, to use comodulation as a segregation cue, spectral resolution must be sufficiently fine to distinguish the time-varying spectral changes of the target speech from the comodulated noise masker. This could account for the SRT difference between modulated speech-shaped noise and steady-state noise in the unprocessed and 24-channel processing conditions. If the spectral resolution is too coarse (e.g., in the 8- and 4-channel conditions) the stimulus representation of target speech will also exhibit very strong comodulation. This may eliminate differences in comodulation as a valid cue for distinguishing between masker and target, and may account for the lack of an SRT difference between the modulated speech-shaped noise and steady-state noise in the 8- and 4-channel processing conditions.

C. Unprocessed versus 24-channel processing

As shown in Fig. 2, performance with 24-channel processing was considerably worse than with no processing, for

all masker types. This may seem surprising, given that the spectral resolution in the 24-channel processing condition was chosen to be similar to that found for normal-hearing listeners, with 3-dB filter bandwidths of 1.16 Cams. In Fig. 4, the excitation patterns (Moore *et al.*, 1997) for the vowel /a/ with and without 24-channel processing are plotted. It can be seen that the spectral peaks of the vowel are comparably well represented in both the processed (dashed) and unprocessed (solid) conditions. This may suggest that the temporal fine structure, discarded by the processing, while not necessary for speech recognition in quiet, may play an important role in segregating speech from interfering sounds. Similarly, it can be seen from Fig. 4 that the spectral resolution of the first few harmonics (below 500 Hz) is degraded in the 24-channel processing condition. While that information may not be important for speech reception *per se*, the first few harmonics carry important information about the stimulus F_0 . It is therefore possible that the loss of F_0 information, due to a reduction in harmonic resolution and/or a loss of temporal fine-structure information, is responsible for the large difference in performance between the unprocessed and 24-channel conditions.

D. Male versus female single-talker interference

As mentioned in Sec. II, the motivation for using different gender single-talker interferers came from the observation that normal-hearing listeners benefit from F_0 differences between target and interfering talkers. This benefit generally increases with increases in F_0 differences (Assmann and Summerfield, 1990; 1994; Bird and Darwin, 1998). Our finding of no significant difference between the male and female interferers may therefore seem surprising. There are at least three possible explanations for this null effect.

The first possible explanation lies in the instantaneous F_0 values. In many past studies of single-talker interference (Brokx and Nooteboom, 1982; Assmann and Summerfield, 1990; 1994; Bird and Darwin, 1998), the F_0 s of the targets and maskers were held constant, either through the use of

short-duration stimuli or synthesized speech with a fixed F_0 . In the present study, the single-talker interferers were taken from recorded books, where exaggerated prosody is common. Although the mean F_0 of the male single-talker interference (111.4 Hz) was approximately equal to that of the target (110.8 Hz), and the mean F_0 of the female single-talker interference (129.4 Hz) was about three semitones higher, the natural variations in F_0 were left unaltered. As a result, the F_0 differences between the target and single-talker interference were distributed such that the probability⁵ of a two-semitone difference in F_0 between the target and male single-talker interference was 0.69, and between the target and the female single-talker interference was 0.76. The lack of difference in SRT values between the male and female masker may therefore be due to the large differences in instantaneous F_0 between the target and both maskers. However, contrary to the hypothesis, previous studies showed little or no improvements in identification as a result of time-varying F_0 s, as compared to constant F_0 s (Darwin and Culling, 1990; Summerfield and Culling, 1992; Assmann, 1999). Their findings suggest that the instantaneous difference in F_0 between the competing talkers in this study may not be the main factor behind the lack of difference between the two single-talker maskers.

The second possible explanation lies in the atypical F_0 range of the female voice used in this experiment. Adult female voices have an average F_0 of around 220 Hz (Hillenbrand *et al.*, 1995). The finding of no significant difference between the male and female interferers in this study may be due to unusually low mean F_0 (129.4 Hz) of the female interferer. The lack of a gender effect in this study may, therefore, not generalize to everyday situations.

The third possible explanation lies in the individual vocal characteristics of the talkers (e.g., vocal tract length, accent, speaking style, sentence level stress, etc.) The differences in vocal characteristics between the target and interfering talkers may have been sufficiently large to render any further improvement due to mean F_0 difference negligible.

E. Importance of frequency selectivity and temporal fine structure: Implications for cochlear implants

Speech is an ecologically important stimulus for humans. If speech reception in quiet can be achieved with minimal spectral resolution and no temporal fine-structure information, then is the exquisite frequency selectivity and sensitivity to fine structure of the human auditory system necessary for speech communication? One important function of frequency selectivity may be found in earlier studies of spectral smearing (ter Keurs *et al.*, 1992; Baer and Moore, 1993) and of cochlear-implant simulations in steady-state noise (Dorman *et al.*, 1998; Fu *et al.*, 1998). From these studies, and from the present study, it can be seen that greater frequency selectivity results in lower signal-to-noise ratios necessary for speech reception. The present results suggest that sensitivity to low-frequency temporal fine structure (e.g., Meddis and O'Mard, 1997), or the spectral resolution of the lower harmonics (e.g., Terhardt, 1974) is critical to good

speech reception in complex backgrounds. Performance is greatly affected by simulated implant processing, even with 24-channel resolution. The effect of stimulus processing is especially dramatic for the single-talker interferers, where an SRT benefit with respect to steady-state noise in the unprocessed condition is transformed into a deficit in all processed conditions. We hypothesize that the dramatic deterioration in performance is related to an inability to perceptually segregate the target from the masker, and that successful segregation relies on good frequency selectivity and F_0 sensitivity.

The present results have possible implications for cochlear-implant design. As in previous studies (Dorman *et al.*, 1998; Fu *et al.*, 1998), the results support separating the spectrum into as many channels as possible, given the technical constraints of ensuring channel independence (Friesen *et al.*, 2001). However, the results also suggest that simply increasing the number of channels (at least to 24) may not assist in providing sufficient F_0 information to successfully segregate a target from an acoustically complex background. The problem of presenting usable fine-structure information to implant users is current topic of research (e.g., Litvak *et al.*, 2001), and the present results provide further support for such endeavors. Finally, the large differences between performance in steady-state noise and performance in fluctuating backgrounds, particularly single-talker interferers, suggest that testing cochlear-implant patients in steady-state noise alone may underestimate the difficulties faced by such listeners in everyday acoustic environments.

V. SUMMARY

- (1) Simulated cochlear-implant processing leads to a large deterioration in speech reception in the presence of a masker, even when the spectral resolution approaches that of normal hearing.
- (2) Under simulated implant processing, single-talker interference is more detrimental to speech reception than steady-state noise. This is the converse of the situation found without processing, and it highlights the potential importance of frequency selectivity and temporal fine-structure information in segregating complex acoustic sources.
- (3) In the presence of steady-state noise, the amplitude modulations associated with the target speech may provide useful source segregation cues, even under stimulated implant processing, provided that the spectral resolution is sufficiently fine.
- (4) Using steady-state noise to test speech intelligibility may underestimate the difficulties experienced by cochlear-implant patients in everyday acoustic backgrounds.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (NIDCD Grant R01 DC 05216). We are grateful to Christophe Micheyl, Jeanie Krause, Peninah Rosengard, and Alan Asbeck for their comments on an earlier version of this manuscript. We also thank Peter Assmann, Bob Shannon, and John Culling for their many helpful suggestions.

- ¹Shannon *et al.* (1995) processed speech by first dividing the audio spectrum into a small number of wide contiguous frequency bands. The temporal envelope of each band was extracted by half-wave rectification and low-pass filtering. The envelope derived from each band was then used to modulate separate white noises. The modulated noise from each envelope was frequency-limited by filtering with the same bandpass filters used in the original analysis band. The resulting modulated noises were then summed together.
- ²This is more frequently referred to as the ERB scale. However, as pointed out by Hartmann (1997), ERB simply refers to equivalent rectangular bandwidth, which could be used to define all estimates of auditory filter bandwidths. We, therefore, follow Hartmann's convention of referring to the scale proposed by Glasberg and Moore as the Cam scale, in recognition of its origins in the Cambridge laboratories. Described in Glasberg and Moore (1990), $Cam = 21.4 \log_{10}(0.00437f + 1)$, where f is frequency in Hz.
- ³Speech reception threshold is the target-to-masker ratio (dB TMR) at which 50% of the words were correctly identified (see Fig. 1).
- ⁴The standard error of fit is the square root of the summed square of error divided by the residual degrees of freedom, $\sqrt{SSE/v}$, where $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. The residual degrees-of-freedom term (v) is defined as the number of response values (n) minus the number of fitted coefficients (m) estimated for the response values, $v = n - m$.
- ⁵The probability of a two-semitone difference in $F0$ was computed by integrating the $F0$ joint probability distribution function of the target and male or female single-talker interference.
- Assmann, P. F. (1999). "Fundamental frequency and the intelligibility of competing voices," Proc. 14th Int. Cong. of Phonetic Sci., pp. 179–182.
- Assmann, P. F., and Summerfield, Q. (1990). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. **88**, 680–697.
- Assmann, P. F., and Summerfield, Q. (1994). "The contribution of waveform interactions to the perception of concurrent vowels," J. Acoust. Soc. Am. **95**, 471–484.
- Baer, T., and Moore, B. C. J. (1993). "Effects of spectral smearing on the intelligibility of sentences in the presence of noise," J. Acoust. Soc. Am. **94**, 1229–1241.
- Bird, J., and Darwin, C. J. (1998). "Effects of a difference in fundamental frequency in separating two sentences," in *Psychophysical and Physiological Advances in Hearing*, edited by A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis (Whurr, London).
- Blamey, P. J., Dowell, R. C., Tong, Y. C., Brown, A. M., Luscombe, S. M., and Clark, G. M. (1984). "Speech processing strategies using an acoustic model of a multiple-channel cochlear implant," J. Acoust. Soc. Am. **76**, 104–110.
- Brox, J. P. L., and Nooteboom, S. G. (1982). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23–36.
- Brungart, D. S. (2001). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.
- Burns, E. M., and Viemeister, N. F. (1976). "Nonspectral pitch," J. Acoust. Soc. Am. **60**, 863–869.
- Burns, E. M., and Viemeister, N. F. (1981). "Played again SAM: Further observations on the pitch of amplitude-modulated noise," J. Acoust. Soc. Am. **70**, 1655–1660.
- Carlyon, R. P., van Wieringen, A., Long, C. J., Deeks, J. M., and Wouters, J. (2002). "Temporal pitch mechanisms in acoustic and electric hearing," J. Acoust. Soc. Am. **112**, 621–633.
- Darwin, C. J., and Carlyon, R. P. (1995). "Auditory grouping," in *Hearing*, edited by B. C. J. Moore (Academic, San Diego).
- Darwin, C. J., and Culling, J. F. (1990). "Speech perception seen through the ear," Speech Commun. **9**, 469–475.
- de Cheveigné, A., and Kawahara, H. (2002). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**, 1917–1930.
- Dorman, M. F., Loizou, P. C., Fitzke, J., and Tu, Z. (1998). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels," J. Acoust. Soc. Am. **104**, 3583–3585.
- Drullman, R. (1995). "Speech intelligibility in noise: Relative contribution of speech elements above and below the noise level," J. Acoust. Soc. Am. **98**, 1796–1798.
- Festen, J. M., and Plomp, R. (1990). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.
- Fletcher, H., and Galt, R. H. (1950). "The perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**, 89–151.
- Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (1999). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.
- Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (2001). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163.
- Fu, Q. J., Shannon, R. V., and Wang, X. S. (1998). "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," J. Acoust. Soc. Am. **104**, 3586–3596.
- Glasberg, B. R., and Moore, B. C. J. (1986). "Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments," J. Acoust. Soc. Am. **79**, 1020–1033.
- Glasberg, B. R., and Moore, B. C. J. (1990). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.
- Glasberg, B. R., and Moore, B. C. J. (1992). "Effects of envelope fluctuations on gap detection," Hear. Res. **64**, 81–92.
- Gustafsson, H. A., and Arlinger, S. D. (1994). "Masking of speech by amplitude-modulated noise," J. Acoust. Soc. Am. **95**, 518–529.
- Hartmann, W. M. (1997). *Signals, Sound, and Sensation* (Springer, New York).
- 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.
- Houtgast, T., Steeneken, H. J. M., and Plomp, R. (1980). "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics," Acustica **46**, 60–72.
- Houtsma, A. J. M., and Smurzynski, J. (1990). "Pitch identification and discrimination for complex tones with many harmonics," J. Acoust. Soc. Am. **87**, 304–310.
- IEEE (1969). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **AU-17**(3), 225–246.
- Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.
- Litvak, L., Delgutte, B., and Eddington, D. (2001). "Auditory nerve fiber responses to electrical stimulation: Modulated and unmodulated pulse trains," J. Acoust. Soc. Am. **110**, 368–379.
- Meddis, R., and O'Mard, L. (1997). "A unitary model of pitch perception," J. Acoust. Soc. Am. **102**, 1811–1820.
- Miller, G. A., and Licklider, J. C. R. (1950). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167–173.
- Moore, B. C. J., Glasberg, B. R., and Baer, T. (1997). "A model for the prediction of thresholds, loudness, and partial loudness," J. Audio Eng. Soc. **45**, 224–240.
- Nilsson, M., Soli, S., and Sullivan, J. (1994). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.
- Oxenham, A. J., and Moore, B. C. J. (1997). "Modeling the effects of peripheral nonlinearity in normal and impaired hearing," in *Modeling Sensorineural Hearing Loss*, edited by W. Jesteadt (Erlbaum, Hillsdale, NJ).
- Patterson, R. D., Nimmo-Smith, I., Weber, D. L., and Milroy, R. (1982). "The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold," J. Acoust. Soc. Am. **72**, 1788–1803.
- Peissig, J., and Kollmeier, B. (1997). "Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners," J. Acoust. Soc. Am. **101**, 1660–1670.
- Peters, R., Moore, B., and Baer, T. (1998). "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," J. Acoust. Soc. Am. **103**, 577–587.
- Plomp, R. (1967). "Pitch of complex tones," J. Acoust. Soc. Am. **41**, 1526–1533.
- Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., and Lang, J. M. (1994). "On the perceptual organization of speech," Psychol. Rev. **101**, 129–156.
- Shackleton, T. M., and Carlyon, R. P. (1994). "The role of resolved and unresolved harmonics in pitch perception and frequency-modulation discrimination," J. Acoust. Soc. Am. **95**, 3529–3540.
- Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M.

- (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- 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. (1998). *Acoustic Phonetics* (MIT Press, Cambridge, MA).
- Summerfield, Q., and Culling, J. F. (1992). "Auditory segregation of competing voices: Absence of effects of FM or AM coherence," *Philos. Trans. R. Soc. London, Ser. B* **336**, 357–365.
- ter Keurs, M., Festen, J. M., and Plomp, R. (1992). "Effect of spectral envelope smearing on speech reception," *J. Acoust. Soc. Am.* **91**, 2872–2880.
- Terhardt, E. (1974). "Pitch, consonance, and harmony," *J. Acoust. Soc. Am.* **55**, 1061–1069.
- Vliegen, J., and Oxenham, A. J. (1999). "Sequential stream segregation in the absence of spectral cues," *J. Acoust. Soc. Am.* **105**, 339–346.
- 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.