Speech, Hearing and Language: work in progress

Volume 11

Effects of the number of channels and speech-to-noise ratio on rate of connected discourse tracking through a simulated cochlear implant speech-processor

Andrew FAULKNER, Stuart ROSEN and Lucy WILKINSON

Department of Phonetics and Linguistics
UNIVERSITY COLLEGE LONDON

Effects of the number of channels and speech-to-noise ratio on rate of connected discourse tracking through a simulated cochlear implant speech-processor

Andrew FAULKNER, Stuart ROSEN and Lucy WILKINSON

Abstract

A number of recent studies have investigated simulations of cochlear implant speech processors with the aim of establishing the minimum number of channels required to support speech perception in quiet and in noise. These studies have all used citation form consonant and vowel stimuli or simple sentences. Intelligibility measures for such materials, especially sentences, can often show ceiling effects. The present study has examined this issue using connected discourse tracking, a task that can be less subject to ceiling effects and is more representative of everyday communication. Speech processing employed a real-time sine-excited vocoder having three, four, eight or 12 channels. Amplitude envelopes extracted from each band modulated sinusoidal carrier signals placed at each band centre frequency. Speech-spectrum shaped random noise was added to speech prior to the vocoder processing to give three signal-to-noise ratios of +7, +12, and +17 dB. Noise levels were adjusted in real time according to measurements of speech level. Connected discourse tracking rates through the vocoders increased significantly with number of channels up to 12 in both quiet and noise, and decreased significantly with each increase in the noise level from quiet. For natural speech, these levels of noise had little effect on tracking rate. We conclude that with connected speech, optimal performance from a cochlear implant in the quiet and in modest levels of noise is likely to require more than eight independent frequency channels.

1. Introduction

A number of recent studies have examined the effect of the number of channels in a simulated CIS cochlear implant on the perception of citation form speech presented in quiet conditions (Dorman, Loizou, & Rainey, 1997; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995) and in noise (Dorman, Loizou, Fitzke, & Zhemin, 1998; Fu, Shannon, & Wang, 1998). These simulations relate not only to cochlear implant speech processing, but also more generally to the effects of noise and the degree of spectral resolution on the auditory processing of speech information. These studies use vocoder processing, in which speech is split into a limited number of frequency channels, each of which is represented by the time-varying amplitude envelope measured over the band. Within-band spectral information is thus discarded, and only time-varying between-band level differences are available to signal spectral structure. Purely temporal cues are preserved up to the modulation rate limit of the extracted envelopes.

With such processing applied to speech in quiet, consonant identification and the intelligibility of words in the relatively simple HINT sentences both show fairly high levels of performance with between four to six spectral bands (Dorman *et al*., 1997; Shannon *et al*., 1995). Vowel identification in these same studies, and the intelligibility of more complex sentences from the TIMIT database (Loizou, Dorman, & Tu, 1999), are only slightly more demanding of spectral detail in the absence of noise. Here, asymptotic scores are found with between six to eight channels. In these studies of speech in quiet it is difficult to distinguish asymptotic performance from

ceiling effects, and hence any effects of limiting spectral resolution may be obscured. For speech in noise, ceiling effects are less evident. Here the intelligibility of HINT sentences at a -2dB speech-to-noise ratio continues to increase with number of channels up to at least eight (Dorman *et al*., 1998). Consonant and vowel identification in noise show a similar pattern, with 16 channels leading to higher scores than 8 channels (Fu *et al*., 1998).

The present study has examined the effects of the number of channels on speech perception in quiet and in noise using Connected Discourse Tracking (CDT: DeFilippo & Scott, 1978). This task allows a measure of communication rate, which is less likely than is a measure of intelligibility to be limited by ceiling effects. The use in CDT of extended and meaningful connected speech also makes it more representative of everyday communication than measures based on isolated citation-form words or single sentences. CDT has recognized limitations both in its inherent non-repeatability and variability, and the possibility that the test talker adapts their speaking level and style to adjust for difficult communication conditions. It nevertheless remains an interpretable measure of performance when robust differences are found and appropriate controls are observed.

2. Method

2.1 Speech Processing

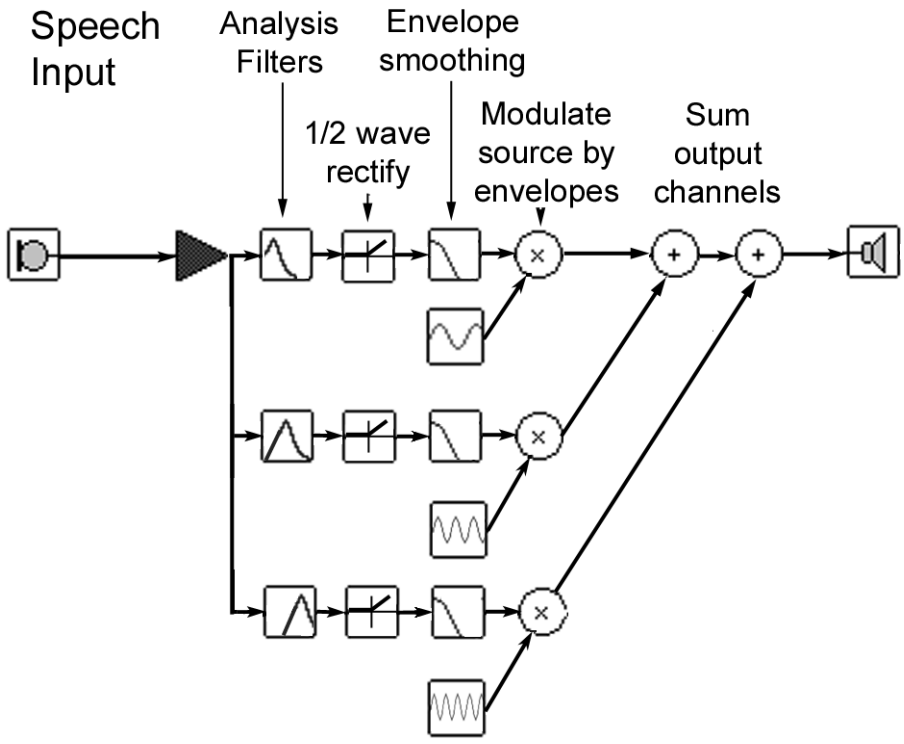

Figure 1 Block diagram of signal processing for a three-channel sine-excited vocoder

Speech processing was carried out in real time using the Aladdin Interactive DSP Workbench (v1.02, AB Nyvalla DSP), and ran at an 11.025 kHz sample rate on a Loughborough Sound Images DSP card with a Texas Instruments TMS320C31 processor. The technique was similar to that described by Dorman *et al*. (1997) in the use of a series of sinusoids as carriers for envelope modulations in each frequency band. The input speech was first low-pass filtered and sampled (16 bits). The signal was then passed through a bank of analysis filters (6th-order elliptical IIR) with frequency responses that crossed 15 dB down from the pass-band peak. Envelope detection occurred at the output of each analysis filter by half-wave rectification and 1st-order low-pass filtering at 30 Hz. These envelopes were then multiplied by sinusoids at the center frequency of the analyzing filter. The modulated sinusoids were summed and played out through a 16 bit digital-to-analogue converter.

Channel Number	3 channel		4 channel		8 channel		12 channel	
	c.f.	-15 dB	**c.f.**	-15 dB	**c.f.**	-15 dB	**c.f.**	-15 dB
		100		100		100		100
1	**234**		**198**		**148**		**132**	
		548		392		219		174
2	**981**		**628**		**293**		**217**	
		755		1005		392		270
3	**2962**		**1519**		**502**		**325**	
		5000		2294		642		392
4			**3387**		**804**		**464**	
				5000		1005		548
5					**1241**		**641**	
						1531		749
6					**1874**		**868**	
						2294		1005
7					**2792**		**1158**	
						3399		1334
8					**4122**		**1530**	
						5000		1755
9							**2006**	
								2294
10							**2616**	
								2984
11							**3397**	
								3868
12							**4398**	
								5000

Table I: Center frequency (c.f.) and -15dB down crossover points of the analysis filters for the 3, 4, 8, and 12 channel processors.

The center frequencies of each analysis filter and the -15 dB crossover frequencies of the filters are shown in Table I. These are based on equal basilar membrane distance

according to the formula given by Greenwood (1990). The frequency of the sinusoidal carrier for each channel was always the same as the analysis filter center frequency.

The real-time processing also controlled the addition of noise. Speech-to-noise ratio was maintained approximately constant through dynamic adaptation to the speech level. The spectral shape of the masking noise was based on the long-term average speech spectrum for male and female voices (Byrne *et al.*, 1994). A close approximation of this spectral shape was produced by filtering a white noise source with a 2nd-order Butterworth bandpass filter. Adaptation of the noise level relative to that of the speech was controlled by a slow-moving amplitude envelope extracted from the speech input. A two-stage process extracted this envelope so that the decay of the envelope in response to speech was slower than its onset. The speech waveform was first full-wave rectified, and then passed through two cascaded 1st-order 1 Hz low-pass filters. This first stage envelope was then further low-pass filtered by a second cascaded pair of 1st-order 1 Hz filters. The envelope used to modulate the noise was the sum of the first stage envelope and the output of the second pair of low-pass filters. The response to an impulse of this envelope extractor is shown in Figure 2. The envelope reaches its maximum 270 ms after the onset of the impulse, and decays to 50% of the maximum 740 ms after onset. In addition, a constant low level noise was present that was 40 dB down from the speech-level related noise component at typical speech input levels.

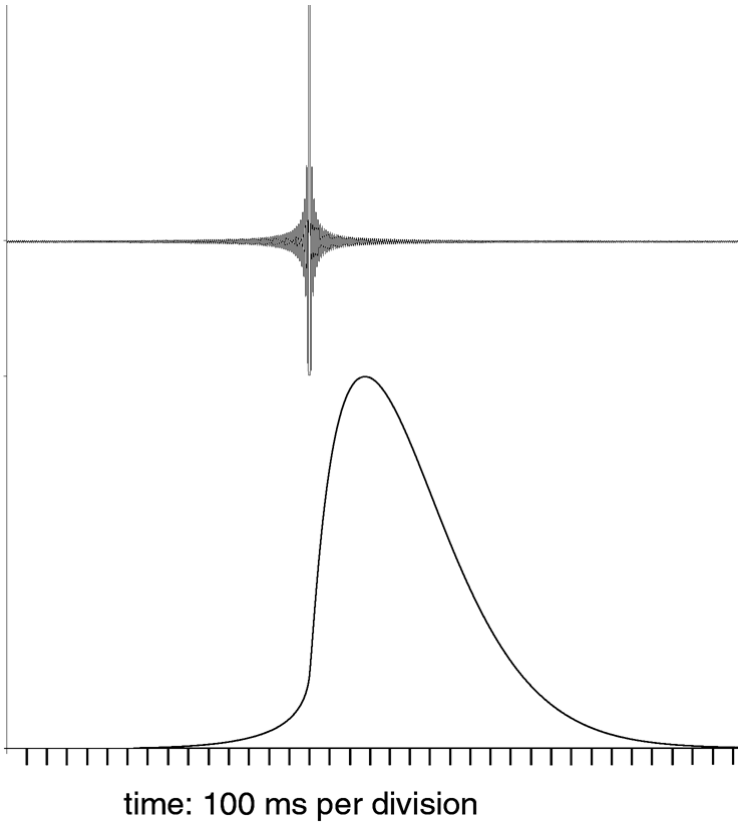

Figure 2: Response of slow envelope extractor (lower panel) to a band-limited impulse (upper panel).

In order to account for the delays imposed by the low-pass filtering, the speech signal was delayed by 3000 samples (272.1 ms) before being added to the noise. This

combined signal was then fed to the system blocks that performed the sine-excited vocoding.

The signal–to-noise ratio (SNR) at the input to the vocoder was calibrated using triggered measurements with a real-time spectral analyzer. Five sentences pre-recorded by the CDT talker were used for these measurements. From pilot testing, SNRs of +17, +12, and +7 dB were selected to cover a range over which the noise caused from mild to more extreme difficulty in CDT.

2.2 Procedure

Texts were chosen from the Heinemann Guided Readers series, elementary level. These texts are controlled in the complexity of content and structure and in vocabulary. The CDT talker (author LW) and the listener sat in adjacent sound-isolated rooms. A constant masking noise at 45 dBA was present in the listener's room in order to mask any unprocessed speech transmitted through the intervening wall. Processed speech was presented diotically to the listener over headphones (Sennheiser HD 475) after amplification (Yamaha P2100). The talker was able to hear the listener's responses over an intercom. The talker read from the text in phrases, and the listener repeated back what she/he had heard. If the listener's response was completely correct, the speaker moved on to the next phrase. Where any word was not correctly repeated, the speaker and listener worked together until the phrase was repeated verbatim. Performance was measured by the average number of words per minute correctly repeated back by the listener during each 5-minute block of CDT.

Four native English-speaking adults having audiometric thresholds within normal limits between 125 and 4000 Hz were paid for their participation. All subjects took part in the unprocessed condition first, in order to familiarize them with the testing procedures. After the first session the four processed speech conditions were presented in a random order according to a Latin square design across subjects. The number of channels of processing was fixed throughout a single session. Each testing session consisted of eight 5-minute blocks of CDT with a short break between blocks. The four noise conditions were presented in turn, in a random order for the first four of these blocks, and repeated in a different order for the second four blocks of the session.

3. Results

Raw CDT rates are presented in Figure 3. A repeated-measures ANOVA was performed using factors of subject, SNR and processing condition - including the unprocessed condition. This showed significant effects ($p<0.001$) of noise level and processor condition. There was a significant interaction between SNR and processing condition. Since this interaction was likely to be due to the lack of an effect of noise for the unprocessed condition, a second repeated-measures ANOVA excluded that condition. This showed only main effects of processing condition [$F(3,9) = 33.3$, $p < 0.001$] and of SNR [$F(3,9) = 106$, $p < 0.001$][i]. An *a priori* contrast test showed that each successive increase in number of channels from three up to 12 led to significant increases in CDT rate. A second *a priori* contrast showed that each increment of noise from quiet to a +7 dB SNR led to a significant decrease in CDT rate.

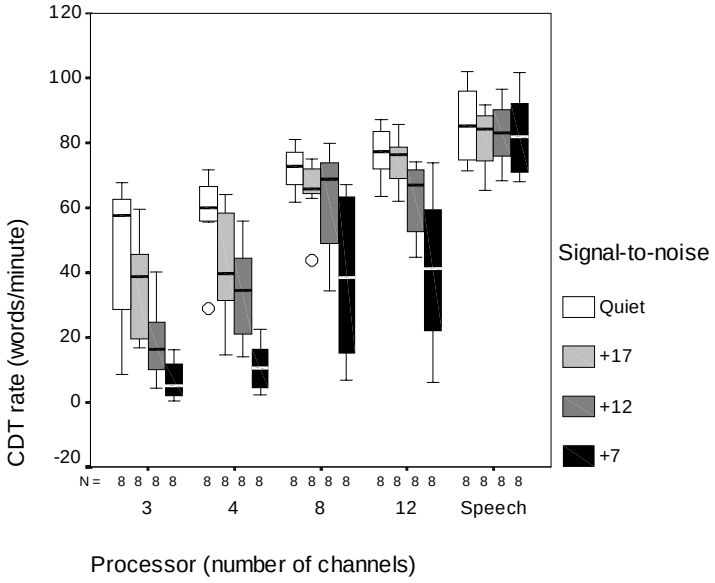

Figure 3: CDT rate as a function of processing condition and SNR. The boxes show the interquartile range, bars within each box represent the median, and whiskers show extreme values. Two outliers, from the same subject, are shown by the symbol ○. These deviate from the median by more than 1.5 times the interquartile range. Each data set contains 8 samples (two CDT sessions from each subject).

The question of whether CDT rates through the processors differed from those with unprocessed speech cannot be addressed from an ANOVA of the whole dataset because of the interaction between processor condition and noise level. Hence, this has been examined through planned contrasts based on sub-analyses in each of the noise conditions to compare each processed condition to the unprocessed speech data. Since unprocessed CDT rates may be at a ceiling level, it is not reasonable to take the lack of a significant difference between these rates and those through a processor as strong evidence for equivalence. However, the presence of a significant difference is readily interpretable. In quiet the unprocessed condition differed only from the three-channel processor. At the +7dB signal to noise ratio, unprocessed speech scores exceeded those from both the three and four channel processors. At +12 dB, unprocessed speech scores significantly exceeded those from the three, four and twelve (but not eight) channel processors. At +7 dB, unprocessed speech scores exceeded those for three and four channels, and were close to being significantly different from the eight channel (p=0.058) and twelve channel (p=0.073) processors.[ii]

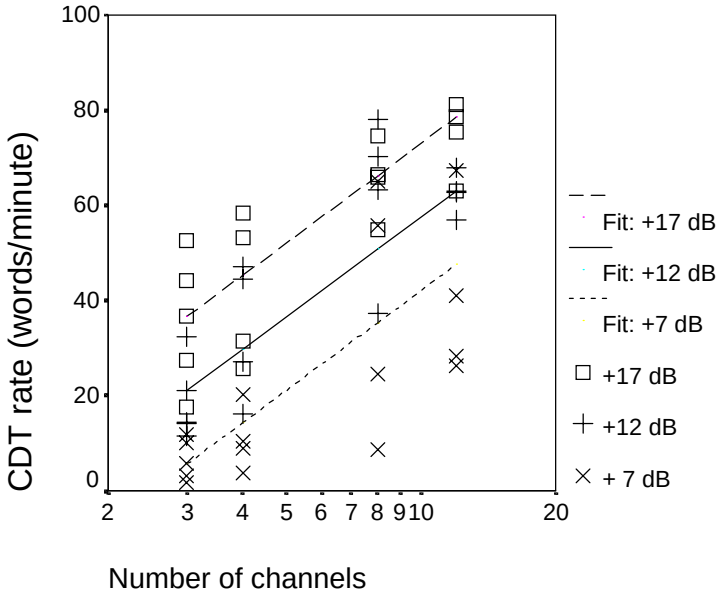

Figure 4: Regression of SNR and log_{10} (number of channels) onto CDT rate for processed speech excluding scores in quiet. The lines are calculated from the regression equation: rate = -49.19 + (SNR dB) x 3.091 + log_{10}(number of channels) x 69.61. Individual symbols are mean CDT rates for the 4 individual subjects at each SNR. The regression accounts for 65.9% of the variance.[iii]

The quantitative effects of number of channels and SNR were estimated through a regression. CDT rate showed a slightly higher correlation with the logarithm of the number of channels than with the untransformed number of channels. The fit of a regression of SNR in dB and $\log_{10}$(number of channels) onto CDT rate is shown in Figure 4. Both the logarithm of the number of channels ($R^2 = 0.423$) and the SNR ($R^2 = 0.239$) contributed significantly to the regression. The overall R^2 of 0.659 is only slightly less than the sums of the squared correlations for the two factors, indicating that the fit is consistent with additive effects of the two variables. The regression indicates that CDT rates increase by about 21 words/minute for each doubling of the number of channels, and by about 18 words/minute for a 6 dB improvement in SNR.

4. Discussion

The CDT rates shown in this study represent measures of communication efficiency with connected speech that are reasonably representative of normal spoken discourse. It is important to establish whether these data are comparable to data from other studies using intelligibility measures for citation form and simple sentence materials.

Figure 5 and Figure 6 show CDT rates together with scores obtained through comparable processing for HINT sentences (Eddington, Rabinowitz, Tierney, Noel, & Whearty, 1997) and consonant and vowel identification (Fu *et al*., 1998). Scores for all measures increase in approximate proportion to the logarithm of the number of channels. There is rather close correspondence between the effects of number of channels and SNR in these data sets. Table II shows correlations between scores for these measures at matching numbers of channels and SNR. The variation of CDT rate with number of channels and SNR correlates significantly with the variation of the other three measures over the SNR range used here (between +7 and +17 dB).

		HINT	VOWELS	CONSONANTS
CDT	r	0.9271	0.8131	0.9879
	p	0.0078 **	0.0013 **	0.0001 **
	N	6	12	12
HINT	r		0.6649	0.9576
	p		0.1032	0.0007 **
	N		7	7
VOWELS	r			0.8533
	p			0.0001 **
	N			29

*Table II: Pearson product-moment correlations between performance in four speech tasks. The data are average scores at each SNR and number of channels, including scores in quiet, but excluding those with unprocessed speech. CDT rates at SNRs of +7 and +17 dB are compared here to performance in other tasks at +6 and +18 dB respectively. ** indicates significance at p<0.01.*

One notable feature of the CDT data is that even a modest level of noise, +17 dB SNR, has a significant impact on scores through the processors. The vowel identification data of Fu *et al*. show similar trends. In the consonant and sentence data, performance at similar SNRs is typically at ceiling levels (see Figure 6). Also striking is that at a relatively moderate SNR of +7 dB, CDT rate with the three and four channel processors has fallen to very low levels of around 5 to 10 words per minute.

In comparable conditions, HINT sentence, vowel and consonant scores are approximately 50% correct. SNRs of the order of +6 dB are relatively common in everyday situations (Pearsons, Bennett, & Fidell, 1977).

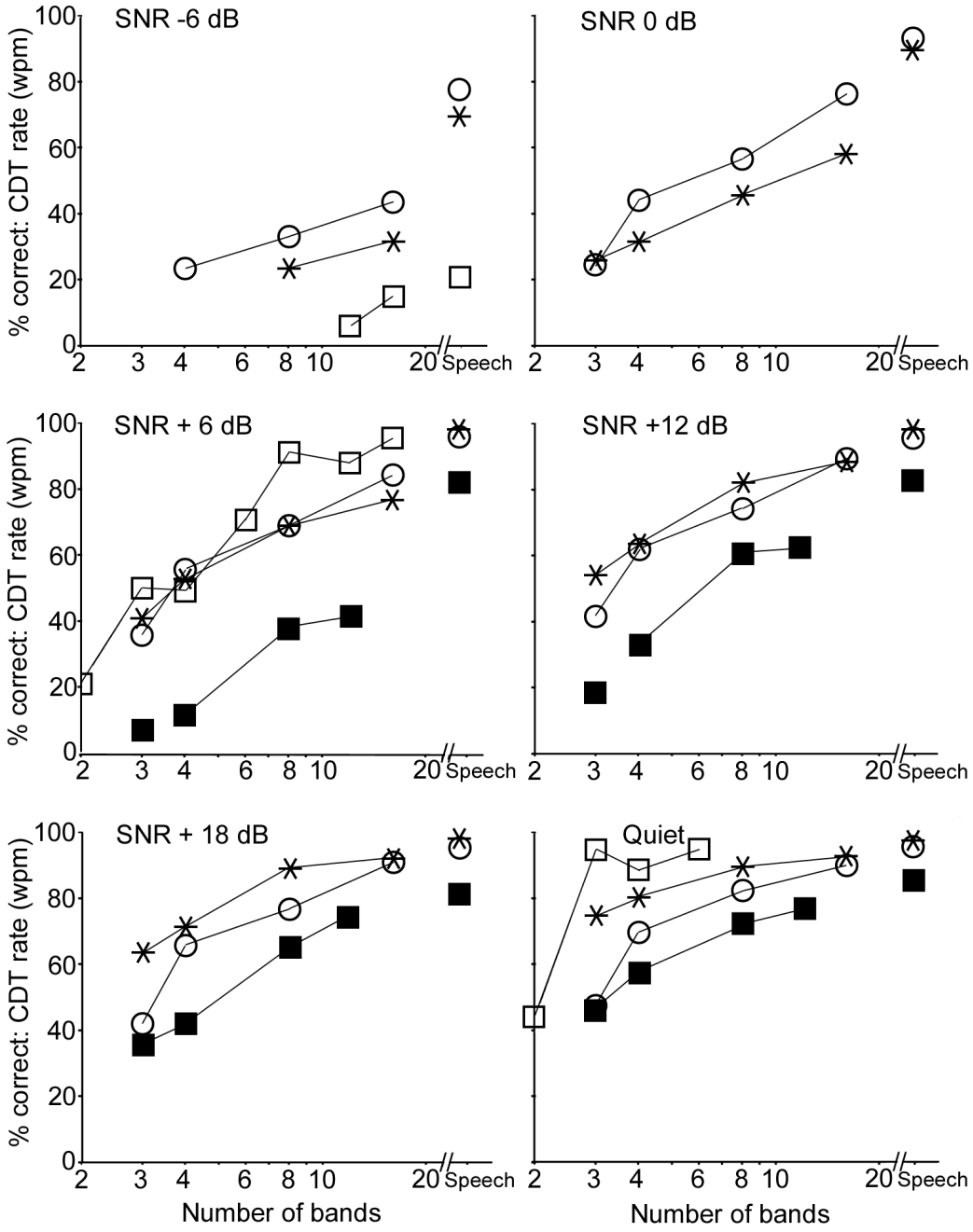

*Figure 5: Effect of channel number on CDT rate, HINT sentence (Eddington et al., 1997), and vowel and consonant identification (Fu et al., 1998). Each panel shows data from a different SNR. Symbols: • CDT rate; • Words correct in HINT sentences; • Vowels correct; * Consonants correct.*

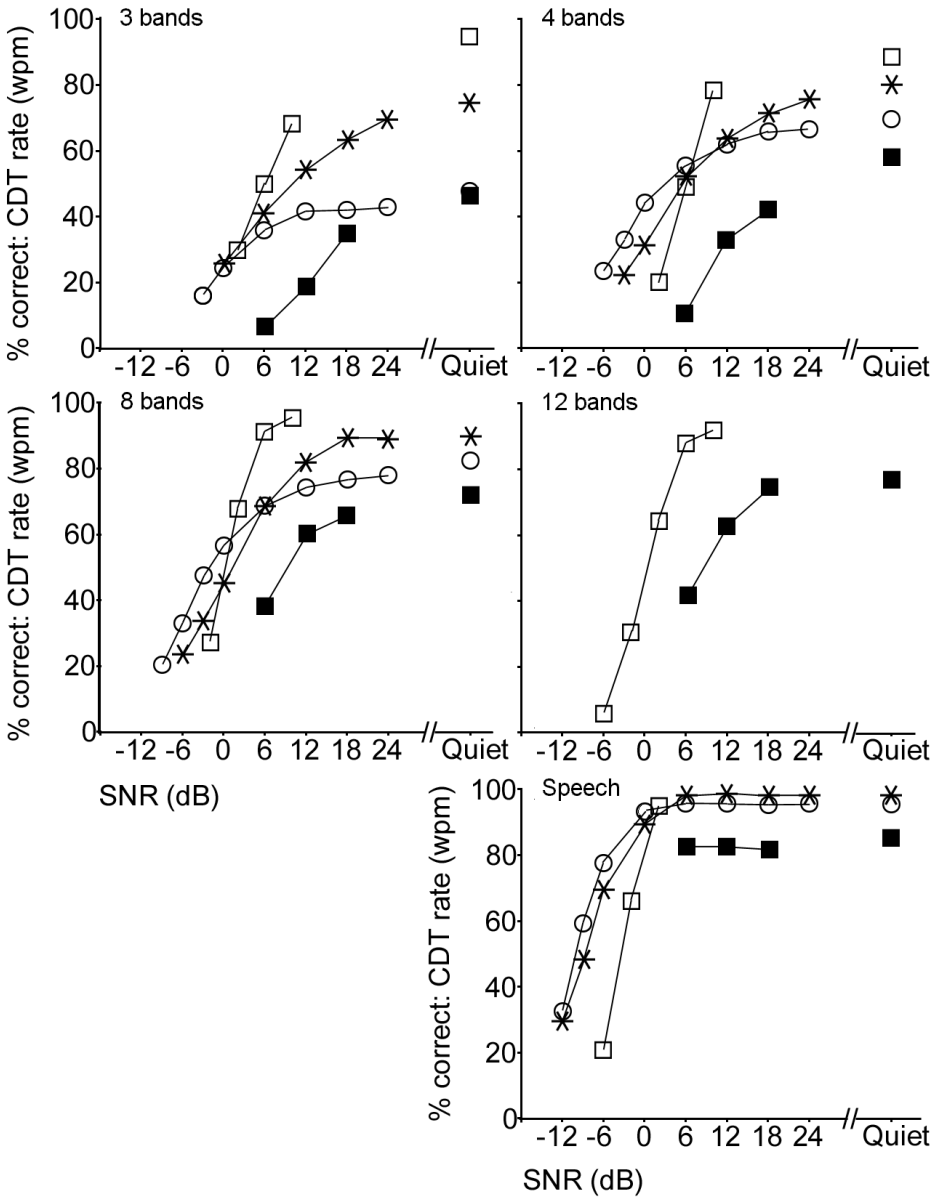

*Figure 6: Effect of SNR on CDT, HINT sentence, vowel and consonant identification. Each panel shows data with a different number of channels. Symbols: • CDT rate; • Words correct in HINT sentences; • Vowels correct; * Consonants correct.*

Other studies have examined the effects of spectral resolution on speech perception using speech processing that smears spectral detail rather than quantizing frequency into a small number of channels as in vocoding. One method of smearing spectral detail involves the computation of amplitude spectra representing the output of a series of band-pass filters. From these smeared amplitude spectra, processed speech is re-synthesized so as to discard spectral detail lost as a result of the limited spectral resolution of the filter-bank. Using such techniques, as with vocoder processors, the intelligibility of sentences in quiet is relatively unaffected even when the filtering uses bandwidths up to six times broader than those of human auditory filters (Baer & Moore, 1993). For comparison with the vocoding results, a set of filters six times broader than the bandwidth of human auditory filters represent between four and five independent frequency channels over a frequency range of 200 to 5000 Hz.

As in the vocoder studies, effects of spectral information loss are much more apparent in noise. At a -3 dB speech-to-noise ratio Baer and Moore found that sentence intelligibility did decline significantly as filter bandwidth was broadened from normal human auditory filter bandwidths to filters three and six times wider. Other studies using comparable methods show similar outcomes (ter Keurs *et al*., 1992; 1993; Leek & Summers, 1996). Baer and Moore found a significant interaction between the degree of smearing and signal-to-noise ratio for sentence intelligibility scores. As in studies of sentence intelligibility through vocoder processors with four to six channels of spectral information (Dorman *et al*., 1998; Eddington *et al*., 1997), sentence scores in quiet were at or close to ceiling levels. A ceiling is not, however, evident for the effects of spectral resolution on Baer and Moore's data in noise, and it seems likely that the interaction they found is a consequence of the ceiling on scores in quiet. CDT rate, as measured here, proves to be free of the difficulties of interpretation that arise from sentence scores close to ceiling levels. CDT rates through the processors used here show no interaction between spectral resolution and signal-to-noise ratio[iv], and reduced spectral resolution has a clear effect both in the quiet and in noise.

5. Summary

The effects of number of channels and noise on CDT rate rather highly correlated with those previously found in sentence identification (Dorman *et al*., 1998; Loizou *et al*., 1999), although much less subject to ceiling effects, and those for in vowel and consonant identification (Fu *et al*., 1998). However, CDT rates are rather low in conditions that give quite high levels of sentence intelligibility.

Across the whole range of SNRs tested, there are significant increases of CDT rate with each increase in number of channels from three to four, to eight, and to 12. Hence, the asymptotic number of channels exceeds eight. For vowel and consonant identification in noise, Fu *et al*. (1998) also found the asymptote to be greater than eight channels, while sentence in noise performance again leads to an estimated asymptote of at least eight channels (Dorman *et al*., 1998). Taken together with the present CDT data, these findings converge on the conclusion that the asymptotic number of channels for speech perception in noise is greater than eight. This is in contrast to the conclusion based on speech in quiet that between four and six channels are sufficient (Dorman *et al*., 1997; Shannon *et al*., 1995).

The addition of noise up to a +7 dB speech-to-noise ratio had no effect on CDT rates for unprocessed speech. For the vocoder processed speech, however, performance was significantly impaired compared to quiet even at a +17 dB speech-to-noise ratio, and declined further with each 5 dB increment in noise level to a +7 dB SNR. With low numbers of channels, which may represent the situation for many cochlear implant users, CDT rates at +7 to +12 dB SNRs were very low. Whereas 50% words correct in sentences in such conditions may seem a reasonably high score in the context of speech intelligibility for a cochlear implant user, these very low CDT rates are perhaps more indicative of the impairment of speech communication that can be expected with a small number of channels in typical noisy environments.

6. Acknowledgements

We are grateful to Don Eddington for permission to reproduce unpublished data. This paper is based on the Lucy Wilkinson's B. Sc. Project.

7. References

Baer, T., & Moore, B. C. J. (1993). "Effects of spectral smearing on the intelligibility of sentences in noise," Journal of the Acoustical Society of America, 94, 1229-1241.

DeFilippo, C. L., & Scott, B. L. (1978). "A method for training and evaluation of the reception of on-going speech," Journal of the Acoustical Society of America, 63, 1186-1192.

Dorman, M. F., Loizou, P. C., Fitzke, J., & Zhemin, T. (1998). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6-20 channels," Journal of the Acoustical Society of America, 104, 3583-3585.

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). "Speech intelligibility as a function of the number of channels for signal processors using sine-wave and noise-band outputs," Journal of the Acoustical Society of America, 102, 2403-2411.

Eddington, D. K., Rabinowitz, W. R., Tierney, J., Noel, V., & Whearty, M. (1997). *Eighth Quarterly Progress Report, October 1, 1997, through December 31, 1997. Speech Processors for Auditory Prostheses. NIH Contract N01-DC-6-2100.* . Cambridge, MA: MIT.

Fu, Q.-J., Shannon, R. V., & Wang, X. (1998). "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," Journal of the Acoustical Society of America, 104, 3586-3596.

Greenwood, D. D. (1990). "A cochlear frequency-position function for several species - 29 years later," Journal of the Acoustical Society of America, 87, 2592-2605.

Loizou, P. C., Dorman, M., & Tu, Z. (1999). "On the number of channels needed to understand speech," Journal of the Acoustical Society of America, 106, 2097-2103.

Pearsons, K. S., Bennett, R. L., & Fidell, S. (1977). *Speech Levels in Various Noise Environments* (EPA-600/1-77-025). Washington, DC: US Environmental Protection Agency.

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

---

[i] Using Huynh-Feldt Epsilon-corrected degrees of freedom where $\varepsilon = 1$ in both cases. The observed power = 1.0 for each test.

[ii] All of these analyses were carried out both with and without the two outlying points. There were no substantive differences in the outcomes.

[iii] The two outliers were excluded from this regression, but once again these affected the analysis minimally.

[iv] This conclusion should be treated with some caution however, as the observed power statistic for this interaction term is only 0.273.