Speech, Hearing and Language: work in progress

Volume 13

Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling (CIS) cochlear implants

Tim GREEN, Andrew FAULKNER and Stuart ROSEN.

Department of Phonetics and Linguistics
UNIVERSITY COLLEGE LONDON

SPECTRAL AND TEMPORAL CUES TO PITCH IN NOISE-EXCITED VOCODER SIMULATIONS OF CONTINUOUS-INTERLEAVED-SAMPLING (CIS) COCHLEAR IMPLANTS

Tim GREEN, Andrew FAULKNER, and Stuart ROSEN.

Abstract

Four-band and single-band noise-excited vocoders were used in acoustic simulations to investigate spectral and temporal cues to melodic pitch in the output of a cochlear implant speech processor. Noise carriers were modulated by amplitude envelopes extracted by half-wave rectification and low-pass filtering at 32 or 400 Hz. The four-band, but not the single-band processors, may preserve spectral correlates of fundamental frequency (*F0*). 400 Hz envelope smoothing preserves temporal correlates of *F0*, which are eliminated with 32 Hz smoothing. Inputs to the processors were sawtooth frequency glides, in which spectral variation is completely determined by *F0*, or synthetic diphthongal vowel glides, whose spectral shape is dominated by varying formant resonances. Normal listeners labelled the direction of pitch movement of the processed stimuli. For processed sawtooth waves, purely temporal cues led to decreasing performance with increasing *F0*. With purely spectral cues, performance was above chance despite the limited spectral resolution of the processors. For processed diphthongs, performance with purely spectral cues was at chance, showing that spectral envelope changes due to formant movement obscured spectral cues to *F0*. Performance with temporal cues was poorer for diphthongs than for sawtooths, with very limited discrimination at higher *F0*. In conclusion, for speech signals through a typical cochlear implant processor, spectral cues to pitch will have little utility, while temporal envelope cues are useful only at low *F0*.

1. Introduction

Voice pitch information plays an important role in the perception of speech, providing cues to linguistic features such as word emphasis and question-statement contrasts (Highnam and Morris, 1987; Nooteboom, 1997; Wells, Peppé, and Vance, 1995), and also to paralinguistic features such as the age, sex, identity and emotional state of the speaker (Abberton and Fourcin, 1978; Lieberman and Michaels, 1962). Both spectral and temporal cues to voice pitch are available to normally-hearing listeners under normal conditions. In continuous interleaved sampling (CIS), a widely used processing strategy for cochlear implants (Wilson *et al.*, 1991), the electrical stimulation delivered to the auditory nerve represents amplitude envelopes extracted from only a small number of spectral bands. These amplitude envelopes are low-pass filtered, typically at 400 Hz, and imposed on biphasic pulse carriers that generally have a fixed rate of 0.8 to 2 kHz. The limited spectral resolution of cochlear implant systems means that the lower harmonics of speech that give normal listeners spectral cues to pitch are not resolved. However, both physiological and psychophysical evidence suggest that the stimulus envelope is unambiguously represented, provided that the carrier pulse rate is 4 to 5 times greater than the maximum modulation frequency (Busby, Tong, and Clark, 1993; McKay, McDermott, and Clark, 1994; Wilson, 1997). Therefore, temporal cues to pitch are, in principle, available in CIS processed speech, as long as the voice fundamental frequency (*F0*) range is passed by the envelope smoothing filter and the pulse rate is high enough.

Evidence to this effect was provided by a study involving users of the LAURA cochlear implant (Geurts and Wouters, 2001). An initial criterion of being sensitive to changes of 20% in the modulation frequency of a sinusoidally amplitude-modulated pulse train for modulation rates around 150 Hz was met by four of eight subjects. The ability of the selected subjects to discriminate differences in the *F0* of synthesized steady-state vowels was compared in conditions in which the low-pass cutoff frequency of the envelope filter was either 400 or 50 Hz. The overall rate of the stimulation was 10,000 pulses per second (pps). Different listeners' implants had either 7 or 8 active channels, so that the pulse rate per channel was either 1250 or 1429 pps. With the 400 Hz envelope filter, performance was relatively good: with a standard *F0* of 150 Hz the difference in *F0* that could be reliably discriminated varied between 6 and 20 Hz across listeners. With the 50 Hz envelope filter, which eliminated temporal envelope fluctuations related to *F0,* performance was generally very much worse. However, there was evidence that stimuli with different *F0*'s could occasionally be discriminated on the basis of loudness cues. Although stimuli were balanced in root mean square level, the average amplitude of the pulses in individual channels was different for different *F0*'s, due to the associated changes in the frequencies of the harmonics of *F0*. For one particular combination of vowel (/i/) and *F0* region (around 250 Hz), it appeared that the pattern of average amplitude changes across channels with *F0* was sufficiently regular to allow reliable discrimination.

In several recent studies CIS processing has been simulated using vocoder-like methods in which temporal envelopes of speech extracted from broad frequency bands modulate noises of the same bandwidths (Dorman, Loizou, and Rainey, 1997; Rosen, Faulkner, and Wilkinson, 1999; Shannon *et al.*, 1995; Shannon, Zeng, and Wygonski, 1998). This processing provides only very limited information on the spectral distribution of energy but maintains temporal envelope information in each band. Provided that there is sufficient envelope bandwidth, such noise-excited vocoding is capable of conveying pitch information for modulation rates up to a few hundred Hz, as indicated by studies using amplitude-modulated noise (e.g., Burns and Viemeister, 1976, 1981; Pollack, 1969). This suggests that the temporal pitch cues available with noise-excited vocoding are broadly similar to those available from CIS processing.

The utility of such temporal pitch cues has thus far received little attention in simulation studies. However, Fu *et al.* (1998) compared the ability of Chinese-speaking listeners to identify Mandarin Chinese's four tonal patterns, characterized by different *F0* contours, under various noise-excited vocoding processing conditions. On each trial listeners were presented with a processed version of a single syllable consisting of an initial consonant and a following vowel with a particular tone, recorded from a single adult male speaker. They were required to identify the tone from a list containing four alternative syllables each of which had the same consonant and vowel. Performance was significantly better when the envelope filter cutoff frequency was 500 Hz rather than 50 Hz, suggesting that listeners were able to make use of the temporal cues to voice pitch that were available when the envelope filter covered the *F0* range. However, performance was well above chance even with the 50 Hz filter, consistent with evidence that cues other than the *F0* pattern, such as amplitude contour and duration, also contribute to tone recognition (Fu and Zeng, 2000; Whalen and Xu, 1992).

Further evidence regarding temporal pitch cues in simulations of CIS processing was provided by Faulkner, Rosen, and Smith (2000). In the context of an investigation of the effects of the salience of pitch and periodicity information on speech intelligibility they constructed a number of four-channel vocoder simulations in which the extent to which such information was conveyed was varied. There were two noise-carrier processors, in which the low-pass cutoff frequency of the envelope filters was either 400 Hz or 32 Hz, the latter eliminating temporal information in the voice pitch range. Other processors used a 32-Hz cut-off frequency, but the selection and control of the carrier signal was employed to convey pitch and periodicity information. For example, in one processor, designated *FxNx*, the carrier during voiced speech was a pulse train the frequency of which followed the fundamental frequency of the speech input, while the carrier during voiceless speech was random noise.

Although speech perception, as assessed in tests of consonant and vowel identification, BKB sentence perception, and connected discourse tracking, was found to differ only slightly across the different processing conditions, it is likely that this is because such tasks lack sensitivity to aspects of speech such as intonation, which nevertheless are likely to be important factors in speech communication. That the processing conditions did differ in the extent to which pitch information was conveyed was confirmed by performance in a frequency glide discrimination task in which listeners were required to categorise as either “rising” or “falling” in pitch sawtooth waves whose fundamental frequency changed smoothly over their 500 ms duration. Performance was very high with the *FxNx* processor, even with small start-to-end frequency ratios, as would be expected since the fundamental frequency is directly represented in the carrier. Performance with the noise-carrier processor with the 400 Hz cutoff envelope filter was substantially worse, being high for large start-to-end frequency ratios but declining to near chance levels for small ratios. Interestingly, performance with the noise processor with the 32 Hz envelope cutoff frequency was only slightly lower than that with the 400 Hz envelope processor despite the elimination of temporal cues to pitch variation. This was attributed to differences in spectral envelope caused by harmonics of the input signal shifting between analysis bands of the processor.

The current study provides a more detailed investigation of the pitch cues available from noise-carrier processors, allowing a dissociation of the contributions of temporal envelope and spectral cues. In our first experiment, sawtooth glide labelling performance was examined with single-band processors with 32 Hz and 400 Hz envelope bandwidths in addition to the two four-band noise-carrier processors described above. With these single-band processors the output was unaffected by the number of harmonics of the sawtooth signal falling in each analysis band, eliminating spectral cues correlated with *F0*. A second experiment examined listeners’ ability to identify the direction of pitch glides when stimuli consisted of synthesized dipthongal vowels, allowing an assessment of the impact of variations in formant structure on cues to voice pitch.

2. Experiment 1: sawtooth glides

2.1 Participants

Seven normally-hearing listeners aged between 25 and 50 years participated, including the three co-authors, three members of departmental staff, and one postgraduate student who was paid £5 per hour.

2.2 Stimuli

Stimuli were generated off-line with a 20 kHz sample rate using MATLAB. They consisted of noise-carrier processed sawtooth glides, constructed from the addition of ten sinusoids, whose fundamental frequency changed logarithmically over their 500 ms duration. The ratio of start and end frequencies varied in six equal logarithmic steps from 1:0.5 to 1:0.93. Three ranges of fundamental frequency were used, with the *F0* at the mid-point in time of each glide being 146, 208, and 292 Hz respectively. For each ratio and each *F0* range there was one ascending and one descending glide, giving a total of 36 different glides.

Four-band noise vocoding comprised the following sequence of steps: analysis bandpass filtering (sixth-order Butterworth) to divide the spectrum into frequency bands; half-wave rectification and low-pass filtering (second-order Butterworth) to extract the amplitude envelope for each band; modulation of a noise carrier by each envelope; output filtering matching the initial analysis filtering; summation across channels. The analysis and output filter bands were based on equal basilar membrane distance (Greenwood, 1990). The filter slopes crossed at their –3 dB cutoff frequencies, which were 100, 392, 1005, 2294, and 5000 Hz. In the single-band conditions processing was identical except that the four analysis filters and envelope extractors were replaced with a 50 Hz high-pass filter followed by a single envelope extractor whose output modulated the level of each of the four output bands. Both four-band and single-band processing were carried out with the cutoff frequency for the low-pass envelope extraction filter at either 32 or 400 Hz, making a total of four processing conditions designated as *Single32, Single400*, *Four32*, and *Four400*.

2.3 Procedure

Stimuli were presented through the right earpiece of Sennheiser HD 414 headphones at a comfortable listening level (peak of 85-90 dB SPL measured over an 80 ms window). On each trial subjects heard a single glide and were required to identify it as either “rising” or “falling” in pitch. They responded via computer mouse by clicking on an image of either a rising or falling line. No feedback was given. Before each block of trials subjects were able to listen to a selection of the glides to be presented in that block, visually labelled as rising or falling. To familiarize themselves with the task subjects were first presented with a block of trials in which the glides were unprocessed. Processing condition was then varied across blocks of trials consisting of three repetitions of the 36 glides in random order. The order in which blocks of trials were presented was random with the constraint that each set of four blocks contained one block with each type of processing. Subjects completed five blocks of trials for each processing condition, the first of which was treated as practice.

2.4 Results and Discussion

Mean psychometric functions for the proportion of "fall" responses as a function of the ratio between start and end frequencies in each processing condition are shown in Figure 1. A logistic regression was carried out on the proportion of "fall" responses as a function of the log (base 10) of the start-to-end frequency ratio for each processing condition and center *F0* for each subject. None of the fits deviated significantly from the observed data according to χ^2-tests with 10 degrees of freedom. Regression intercept values indicated that most listeners showed a slight bias towards "rise" responses, indicated by negative intercepts, in the *Single32* condition. In the other three conditions intercepts varied substantially across listeners, though several listeners showed a bias towards "fall" responses for glides in the lowest *F0* range and for a bias towards "rise" responses for those in the highest *F0* range.

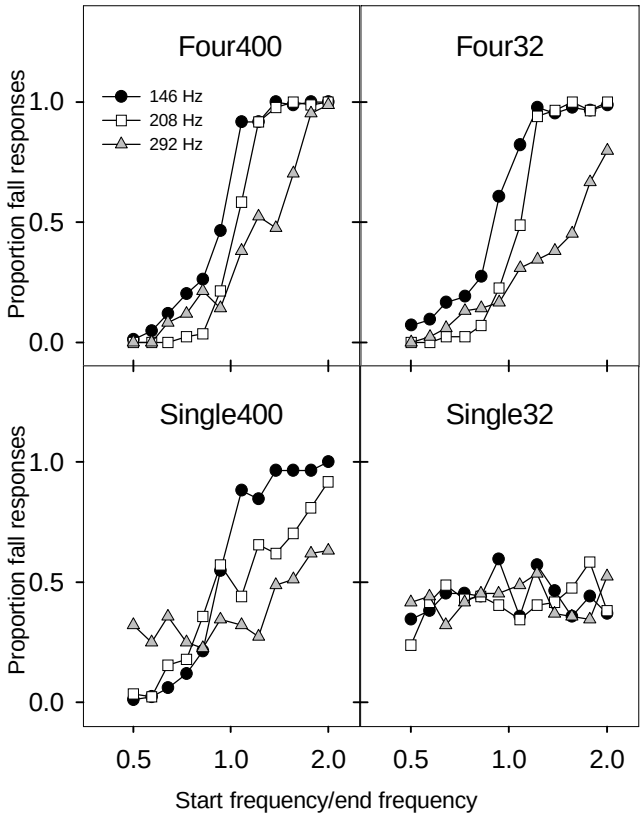

Figure 1*: Proportion of "fall" responses summed across listeners as a function of start-to-end frequency ratio for each processing condition of Experiment 1.*

Figure 2 shows the regression slope estimates, with a larger value indicating greater discriminability. For example, for unbiased responding, a slope of 10 corresponds to a proportion of "fall" responses of 0.58 for the falling glide with the smallest frequency ratio (0.93), while for a slope of 30 the corresponding proportion would be 0.72. A 95% confidence interval was calculated for each slope estimate, but for clarity these

are not displayed in Figure 2. Confidence intervals increased in size with increasing slope. For example, for a slope value of 10 the confidence interval was typically around +/- 3, while for a slope of 30 it was around +/-12. Slope estimates were analysed using a two-way repeated-measures analysis of variance (ANOVA) with factors of processing condition and center *F0.* Since higher slope estimates have a higher standard error, data were logarithmically transformed before analysis. To ensure that there were no negative values, 2 was added to all values before the (base 10) logarithms were taken. The reported *F* tests used Huynh-Feldt epsilon correction factors. There were significant effects of processing condition [F (3,18) = 178.60, $p < 0.001$] and center *F0* [F (2,12) = 25.08, $p < 0.001$], and also a significant interaction between these two factors [F (6,32) = 17.29, $p < 0.001$].

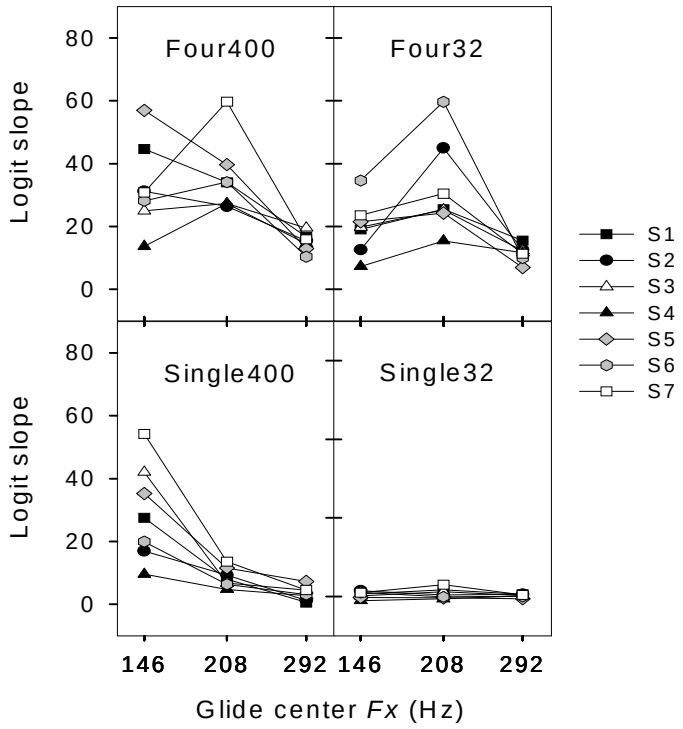

Figure 2*: Slopes of logistic regressions of the proportion of "fall" responses as a function of the log (base 10) of start-to-end frequency ratio for each listener in each processing condition of Experiment 1.*

The nature of these effects and their interaction can be interpreted from the patterns of performance shown in Figures 1 and 2. In the *Single32* condition listeners were clearly unable to distinguish between rising and falling glides. For each *F0* range the mean proportion of "fall" responses is effectively constant regardless of start-to-end frequency ratio (Figure 1), and the slopes of the logistic regressions are virtually zero for all listeners and *F0* ranges (Figure 2). This is consistent with the expectation that neither spectral nor temporal cues to pitch are available with this form of processing. In the S*ingle400* condition, in which there are temporal envelope cues to pitch but no

spectral information, performance varies substantially according to *F0* range. For the lowest *F0* range, the slope of the mean psychometric function is quite steep and performance is near ceiling for several of the larger frequency ratios (Figure 1), indicating that listeners found it relatively easy to distinguish between rising and falling glides. However, the functions become less steep with increasing glide center *F0*. For the highest *F0* range there is only a small change in the proportion of "fall" responses according to start-to-end frequency ratio and performance is well below ceiling for even the largest ratios. The logistic regressions for individual listeners (Figure 2) all show a monotonic decrease in slope estimates as glide center *F0* increases, although there is substantial variation in slope values for the lowest *F0* range. These data illustrate the declining utility of temporal envelope cues to pitch as *F0* increases and also suggest that there is considerable variability in listeners' ability to make use of such cues at lower *F0* values.

The *Four32* condition isolated the contribution of spectral cues in the absence of temporal envelope information. Performance in this condition again varied according to *F0* range, but in a markedly different way to that apparent in the *Single400* condition. The mean psychometric functions (Figure 1) indicate that the highest discriminability occurred with the 208 Hz center *F0*, and the lowest with the 292 Hz center *F0*. For each individual listener the highest regression slope value was observed for the middle *F0* range (Figure 2). For five out of seven listeners slopes were greater for the low *F0* range than the high *F0* range, though there is considerable variability in slope estimates across listeners for the two lower center *F0* values. Pitch information, encoded as spectral differences that arise as harmonics of the glide stimuli move between analysis bands, is clearly available for all three *F0* ranges tested. That this information is more salient with a glide center *F0* of 208 Hz may be attributable to the fact that the second harmonic of stimuli at this center *F0* will be changing between values centered at 416 Hz, close to the point at which the slopes of the first two analysis filters cross (392 Hz).

In the *Four400* condition both spectral and temporal envelope cues to pitch variation were available. The mean psychometric functions (Figure 1) are very similar to those obtained in the *Four32* condition for the two lower *F0* ranges, while a slightly steeper slope is apparent with the 292 Hz center *F0*. In contrast to the *Single400* and *Four32* conditions, the slope values for individual listeners (Figure 2) do not conform to a single pattern. For three listeners (S1, S5, S7) the pattern of slope values is similar to that in the *Single400* condition, with slope values declining consistently as glide center *F0* increases, while for the other four listeners the pattern is similar to that in the *Four32* condition, with the highest slope value occurring with the 208 Hz glide center *F0*. The close similarity in the mean psychometric functions for 146 Hz center *F0* stimuli across these three conditions would suggest that there is little integration of spectral and temporal sources of pitch information. However, the variation in the pattern of slope values across glide center *F0* in the *Four400* condition indicate that there are considerable individual differences in the ways in which temporal and spectral cues are combined and weighted.

Experiment 1 has demonstrated that, in addition to the temporal envelope cues available at low modulation rates, noise-excited vocoding can also provide spectral cues to pitch due to the movement of harmonics between the processor's analysis bands. However, it appears likely that when a series of carrier bands are modulated by envelopes extracted from speech, rather than from a steady-state periodic sawtooth

waveform, spectral cues to pitch are likely to be diminished by the time-varying spectral envelope of speech. This is investigated in Experiment 2.

3. Experiment 2: synthesized diphthongs

3.1 Stimuli

Versions of the diphthongs /aʊ/, /eɪ/, /aɪ/, and /oɪ/ with a duration of 620 ms were created using an implementation of the KLSYN88 Klatt synthesizer with a sample rate of 20 kHz and parameters specified every 5 ms. Formant values were estimated by examining recordings of each diphthong embedded in a cVc context spoken by a male Southern British English speaker. The formant frequency trajectories for each diphthong are portrayed in Figure 3. For each diphthong, *F0* was varied in the same way as in Experiment 1, i.e., in each of three *F0* ranges there were six rising and six falling versions, making a total of 144 different stimuli. The start-to-end frequency ratios and center *F0* values used were identical to those in Experiment 1. The same formant values were used regardless of the variation in *F0*. This differs from real speech, in which a higher *F0* is typically accompanied by higher formant frequencies. In addition, because the first formant frequency (*F1*) of three of the four synthesized vowels decreases towards around 300 Hz (Figure 3), *F0* was close to or higher than *F1* towards the end of the duration of several of the stimuli in the highest *F0* range. Nonetheless, this simplified stimulus set was expected to address adequately the central question of the effects of speech-like spectral variation on spectral and temporal pitch cues. The *Single32* condition was omitted, but in all other respects noise-excited vocoder processing was identical to Experiment 1.

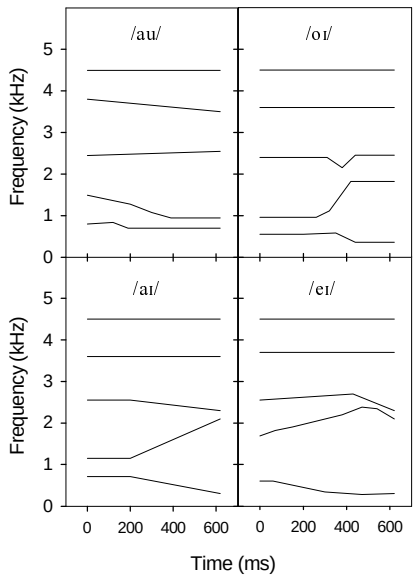

Figure 3: *Trajectories of the formant frequencies of each of the synthesized diphthongs used in Experiment 2.*

3.2 Subjects and procedure

Six of the seven subjects from Experiment 1 completed seven blocks of trials for each processing condition, with the first block regarded as practice. Each block consisted of one presentation of each of the 144 combinations of diphthong, *F0* range, and start-to-end frequency ratio. Other details were as in Experiment 1.

3.3 Results and Discussion

Figure 4 shows the proportion of "fall" responses averaged across the six listeners and the four diphthongs as a function of the ratio between start and end frequencies for each *F0* range and processing condition. Logistic regressions were carried out as in Experiment 1 and again, none of the fits deviated significantly from the observed data according to χ^2-tests with 10 degrees of freedom. Intercept values showed substantially less variation than in Experiment 1 and generally indicated largely unbiased responding, except in the *Four400* condition where most listeners showed a small bias towards "fall" responses for the lowest *F0* range and towards "rise" responses for the highest *F0* range. Regression slope estimates, displayed in Figure 5, were transformed and submitted to an ANOVA as in Experiment 1. There were significant effects of processing condition [F (2,10) = 33.52, $p < 0.001$] and center *F0* [F (2,10) = 14.74, $p = 0.001$], and also a significant interaction between these two factors [F (4,20) = 6.17, $p = 0.002$].

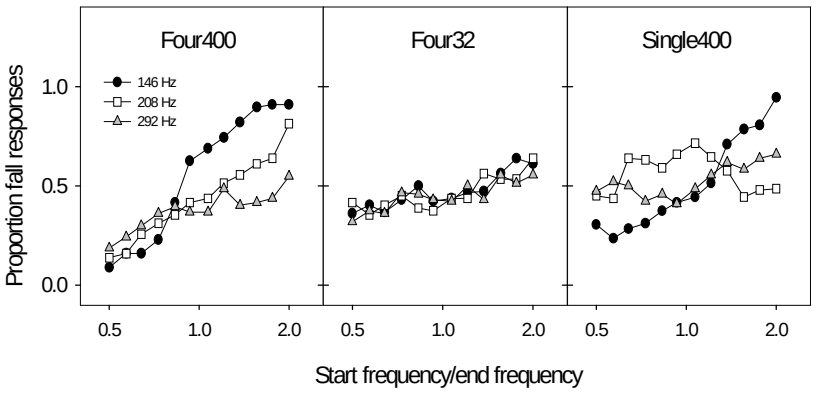

Figure 4: *Proportion of "fall" responses summed across listeners as a function of start-to-end frequency ratio for each processing condition of Experiment 2.*

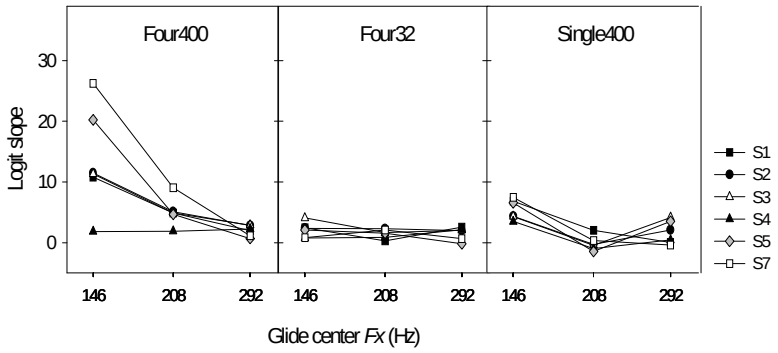

Figure 5*: Slopes of logistic regressions of the proportion of "fall" responses as a function of the log (base 10) of start-to-end frequency ratio for each listener in each processing condition of Experiment 2.*

For all three *F0* ranges, mean performance in the *Four32* condition was barely above chance levels even at the largest frequency ratios, and the slopes of the logistic regressions are close to zero for all listeners and *F0* ranges. This indicates that the variations in spectral envelope associated with diphthongal stimuli obscured the spectral cues to pitch that listeners were able to utilise effectively under the same processing conditions with the sawtooth glides used in Experiment 1. Figure 6 shows mean psychometric functions for each of the four diphthongs for each processing condition and center *F0*. Consistent with the above argument, it is noticeable that responses in the *Four32* condition appear to be largely determined by the identity of the vowel rather than the start-to-end frequency ratio: /aʊ/'s were generally perceived as falling in pitch and /eɪ/'s and /oɪ/'s as rising in pitch, regardless of the actual direction of *F0* change

Comparison of performance across Experiments 1 and 2 is complicated somewhat by the fact that the sawtooth waveform stimuli in Experiment 1 were of a shorter duration (500 ms) than the synthesized diphthongs used in Experiment 2 (620 ms), so that *F0* was changing slightly more quickly in the former case. However, it seems unlikely that this small difference in stimulus duration would have more than a minor effect on performance. This is supported by the results of extra blocks of trials carried out by subject 3 using sawtooth waveform glides of 620 ms duration in processing conditions *Four400*, *Four32*, and *Single400*. Slopes of logistic regressions carried out on these data differed, according to 95% confidence intervals, from the results obtained with the 500 ms glides of Experiment 1 in only two cases (the two higher center *F0* ranges in the *Single400* condition). In both of these cases slope values were actually higher with the 620 ms glides suggesting that the longer duration of the stimuli in Experiment 2 does not contribute to the poorer performance relative to Experiment 1.

In the *Single400* condition mean performance is at or very near chance levels for the two higher *F0* ranges, while there is some limited pitch information available in the lowest *F0* range. Regression slopes vary little across listeners and are in line with the pattern of the mean data, being at or near zero for the two highest *F0* ranges and just above zero for the lowest center *F0*. This is in contrast to the results of Experiment 1, in which substantially steeper slopes were apparent for the lowest center *F0*. This reduction in performance can be attributed to the more complex temporal envelope structure associated with diphthongal stimuli compared to sawtooth waveforms.

Because of the more variable spectral shape of the diphthongal stimuli, *F0*-related fluctuations in the extracted amplitude envelope are less distinct than is the case with sawtooths.

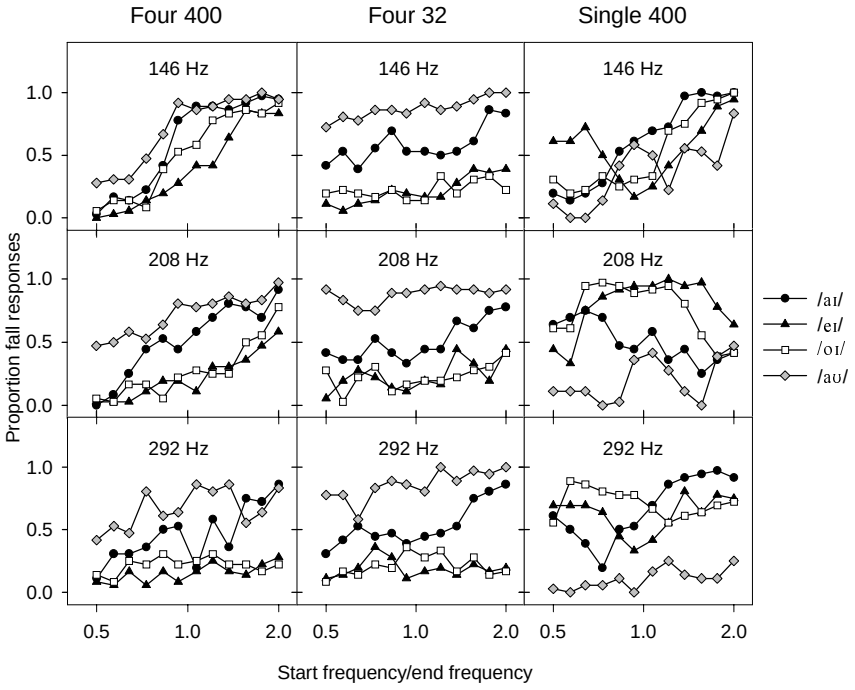

Figure 6*: Proportion of "fall" responses summed across listeners as a function of start-to-end frequency ratio for each vowel in each processing condition of Experiment 2.*

In the *Four400* condition the slopes of the mean psychometric functions decreased monotonically with increasing center *F0*. With the exception of subject 4, whose performance was at or near chance for all *F0* ranges, individual listeners' regression slope values follow a similar pattern, being near zero for the 292 Hz center *F0*, slightly above zero for the 208 Hz center *F0*, and substantially above zero for the 146 Hz center *F0*. Comparing performance in the *Four400* and *Single400* conditions it appears that, for the limited range of *F0* values in which temporal cues to pitch are available, such cues are more effective when envelopes are extracted from four separate channels rather than a single broad bandwidth channel. It is possible that this advantage for the four-band condition is attributable to information carried in the envelope extracted from the lowest frequency channel, the cutoff frequencies of which (100-392 Hz) cover the *F0* range.

4. General discussion and conclusions

4.1 Pitch cues in noise-excited vocoding

Despite the very limited spectral resolution and the use of a 32 Hz cutoff envelope filter, which removed *F0*-related temporal fluctuations, relatively good glide-labelling performance was obtained in the *Four32* condition of Experiment 1. This confirms the

conclusion of Faulkner *et al*. (2000) that with sawtooth waveform stimuli, spectral envelope shifts arising from harmonics of the input waveform moving between analysis bands provide cues to pitch variation. However, the chance levels of performance apparent with synthesized diphthongal stimuli in the *Four32* condition in Experiment 2 demonstrate that such spectral cues are obscured by the presence of variations in spectral structure typical of speech. Therefore, with speech stimuli, reliable perception of pitch variation in noise-excited vocoding simulations of CIS processing will primarily depend upon temporal cues.

The current data indicate that the effectiveness of temporal cues is limited to the lower end of the typical range of voice *F0*. When spectral cues were eliminated by using single-band processing or when input consisted of synthesized diphthongs, discrimination of pitch variation was severely limited for stimuli in the two higher *F0* ranges. In addition, comparison across the *Single400* conditions of Experiments 1 and 2 suggests that the utility of temporal cues to pitch is much reduced when stimuli have a more complex temporal envelope typical of spectrally varying speech. With such stimuli only very limited pitch discrimination was possible in the *Single400* condition. For all but one listener, the utility of temporal cues with diphthongal stimuli in the low *F0* range was substantially higher with four bands rather than a single band. At first sight this result is in contrast to the finding reported by Fu *et al*. (1998) that recognition of noise-excited vocoder processed Mandarin tones, averaged across tones, did not differ according to the number of analysis bands (1, 2, 3, or 4). In that study, however, the distribution of the different tones varied across different analysis band conditions (Fu, 2001, personal communication). The extent to which amplitude contour is correlated to *F0* pattern, and therefore provides a cue to tone recognition in the absence of pitch information, differs markedly across tones. Therefore, this unequal distribution of tones might have obscured an effect of the number of analysis bands. Indeed, this is suggested by Fu *et al*.'s (1998) Figure 2, which presents recognition performance for individual tones in the one band and four band conditions with 500 Hz envelope filtering, averaged across those subjects who also completed a sentence recognition task. For three of the four tones performance was substantially higher with four bands rather than one, consistent with the better performance in the *Four400* relative to the *Single400* condition observed with the synthesized diphthongal stimuli in the present study. However, other unpublished tone recognition data (Fu, 2001, personal communication) obtained in conditions in which the distribution of different tones was consistent across various noise-carrier processing conditions shows only a very slight improvement in performance for four bands relative to a single band. In this case higher numbers of analysis bands were required to achieve substantial improvements in tone recognition relative to single band conditions.

One important factor in determining the extent to which voice pitch perception is improved by an increase in the number of analysis bands of a noise-excited vocoder simulation of CIS processing is likely to be the frequency range covered by the analysis bands. In the four band conditions of Fu *et al*. (1998) the corner frequencies of the analysis bands used were 100, 800, 1500, 2500, and 4000 Hz. In the current study analysis bands were based on equal basilar membrane distance resulting in corner frequencies of 100, 392, 1005, 2294 and 5000 Hz. In the unpublished study referred to above analysis bands were also based on equal basilar membrane distance but covered the range between 300 and 6000 Hz. Both the lowest frequency covered and the widths of the bands are likely to influence the extent to which voice pitch

information can be derived. The current results suggest that cues to voice pitch variation in noise-carrier simulations in the presence of spectral variation may be more salient when there is an analysis band covering a relatively narrow range of frequencies encompassing the *F0* values present in the stimulation.

4.2 Implications for cochlear implants

In determining the implications of the current data for speech processing strategies for cochlear implants a limitation of noise-excited vocoder simulations must be acknowledged. Because the carrier in a CIS processor is a high rate stream of pulses rather than random noise, the *F0*-related modulation of the carrier is noise-free. Therefore, neural responses to this pulse-carrier stimulation are likely to be more strongly synchronized to the modulation (Wilson, 1997), so it is possible that pitch information derived from temporal cues will be more salient in CIS processing than in noise-carrier simulations. Note though, that the temporal complexity of envelopes derived from typical speech suggests that the neural firing patterns resulting from electrical stimulation are unlikely to represent clearly the period of *F0*. Only very limited information is available regarding implant users' perception of pitch variation signalled by modulation of the amplitude of pulse trains, as in CIS processing. In addition to the data obtained by Geurts and Wouters (2001) using synthesized vowel stimuli, the pitch percepts elicited by sinusoidally amplitude modulated pulse trains have been studied in selected subjects (McDermott and McKay, 1997; Wilson *et al.*, 1997). There is, though, a lack of information on typical performance over a range of modulation rates. There must, therefore, be some uncertainty regarding the extent to which performance in noise-carrier simulation studies accurately reflects the likely performance of implant users. However, on the assumption that simulation data gives at least a reasonably accurate indication of the temporal pitch cues available to implant users, the current results prompt important considerations regarding the implementation of CIS processing strategies.

It would appear that with current processing methods the temporal cues on which implant users will depend for information on voice pitch variation are severely limited for fundamental frequencies above 200 Hz, and that even for lower *F0* values, their utility is substantially reduced in the presence of spectral variation typical of speech. Given the important contribution of pitch information to speech understanding in everyday situations, methods of processing that enhance the availability of temporal pitch cues would be expected to be of substantial benefit to implant users, though as Faulkner *et al.* (2000) note, more sensitive speech perception tests than those generally used currently may be required in order to demonstrate such benefits in the laboratory. Geurts and Wouters (2001) devised one CIS processing algorithm designed to enhance the availability of pitch information relative to standard CIS processing. One feature of this algorithm was the elimination of phase distortion from the analysis filters, so that the maxima of *F0*-related fluctuations in each channel of the pulsatile output coincided in time. A second was the use of two envelope filters with cut-off frequencies of 400 Hz and 50 Hz. The output of the 50 Hz filter was subtracted from the output of the 400 Hz filter in order to increase the modulation depth of *F0*-related fluctuations. However, there were no significant differences in the ability to discriminate changes in the *F0* of synthesized steady-state vowels processed with this algorithm relative to standard CIS processing. Other strategies for enhancing temporal pitch cues need to be investigated.

That temporal cues appear to be restricted to lower *F0* ranges is of particular concern given the increasing prevalence of implantation in very young children. Intonation is widely held to play an important role in early language development (Jusczyk, 1997), and is markedly exaggerated in child-directed speech (Fernald *et al*., 1989). However, the voice pitch range of young children covers around 250 to 400 Hz, while the female voice pitch range typically extends well above 200 Hz. Therefore, with current processing strategies, implanted young children will be unable to perceive much of the pitch information in their own speech and in other speech to which they are exposed during development. Substantial benefits to language development might be expected to result from the availability of information concerning pitch variation over a higher *F0* range. Since it is the pattern of pitch change that carries most of the relevant information, rather than absolute pitch values per se, one potential solution might involve lowering the rate of *F0*-related modulations. A rate-lowering approach has been shown to be effective in the context of an aid to lip-reading consisting of single electrode external to the cochlea that provided *F0* information (Fourcin *et al*., 1984).

5. Acknowledgements

This study was supported by the Royal National Institute for Deaf People (UK).

6. References

Abberton, E., & Fourcin, A. (1978). Intonation and speaker identification. Language and Speech, 21, 305-318.

Burns, E. M., & Viemeister, N. F. (1976). Nonspectral pitch. Journal of the Acoustical Society of America, 60, 863-869.

Burns, E. M., & Viemeister, N. F. (1981). Played-again SAM: Further observations on the pitch of amplitude-modulated noise. Journal of the Acoustical Society of America, 70, 1655-1660.

Busby, P. A., Tong, Y. C., & Clark, G. M. (1993). The perception of temporal modulations by cochlear implant patients. Journal of the Acoustical Society of America, 94, 124-131.

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

Faulkner, A., Rosen, S., & Smith, C. (2000). Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants. Journal of the Acoustical Society of America, 108, 1877-1887.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fului, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. Journal of Child Language, 16, 477-501.

Fourcin, A., Douek, E., Moore, B., Abberton, E., Rosen, S., & Walliker, J. (1984). Speech pattern element stimulation in electrical hearing. Archives of Otolaryngology, 110, 145-153.

Fu, Q.-J., & Zeng, F.-G. (2000). Identification of temporal envelope cues in Chinese tone recognition. Asia Pacific Journal of Speech, Language and Hearing, 5, 45-57.

Fu, Q.-J., Zeng, F.-G., Shannon, R. V., & Soli, S. D. (1998). Importance of tonal envelope cues in Chinese speech recognition. Journal of the Acoustical Society of America, 104, 505-510.

Geurts, L., & Wouters, J. (2001). Coding of the fundamental frequency in continuous interleaved sampling processors for cochlear implants. Journal of the Acoustical Society of America, 109, 713-726.

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.

Highnam, C., & Morris, V. (1987). Linguistic stress judgements of language learning disabled students. Journal of Communication Disorders, 20, 93-103.

Jusczyk, P. (1997). *The Discovery of Spoken Language*. Cambridge, MA: MIT Press.

Lieberman, P., & Michaels, S. B., (1962). Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. Journal of the Acoustical Society of America, 34, 922-927.

McDermott, H. J. , & McKay, C. M. (1997). Musical perception with electrical stimulation of the cochlea. Journal of the Acoustical Society of America, 101, 1622-1631.

McKay, C. M., McDermott, H. J., & Clark, G. M. (1994). Pitch percepts associated with amplitude-modulated current pulse trains by cochlear implantees. Journal of the Acoustical Society of America, 96, 2664-2673.

Nooteboom, S. (1997). The prosody of speech: Melody and rhythm. In W. J. Hardcastle & J. Laver (Eds.), *The Handbook of Phonetic Sciences.* (pp. 640-673). Oxford:Blackwell.

Pollack, I. (1969). Periodicity pitch for white noise – fact or artefact. Journal of the Acoustical Society of America, 45, 237-238.

Rosen, S., Faulkner, A., & Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. Journal of the Acoustical Society of America, 106, 3629-3636.

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

Shannon, R. V., Zeng, F.-G., & Wygonski, J., (1998). Speech recognition with altered spectral distribution of envelope cues. Journal of the Acoustical Society of America, 104, 2467-2476.

Wells, B., Peppé, S., & Vance, M. (1995). Linguistic assessment of prosody. In K. Grundy (Ed.), *Linguistics in Clinical Practice*. London :Whurr.

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin Tones in the amplitude contour and in brief segments. Phonetica, 49, 25-47.

Wilson, B., Finley, C., Lawson, D., Wolford, R., Eddington, D., & Rabinowitz, W. (1991). Better speech recognition with cochlear implants. Nature, 352, 236-238.

Wilson, B. (1997). The future of cochlear implants. British Journal of Audiology, 31, 205-225.

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.