Effects of Envelope-Vocoder Processing on F0 Discrimination and Concurrent-Vowel Identification

Michael K. Qin and Andrew J. Oxenham

Objective: **The aim of this study was to examine the effects of envelope-vocoder sound processing on listeners' ability to discriminate changes in fundamental frequency (F0) in anechoic and reverberant conditions and on their ability to identify concurrent vowels based on differences in F0.**

Design: **In the first experiment, F0 difference limens (F0DLs) were measured as a function of number of envelope-vocoder frequency channels (1, 4, 8, 24, and 40 channels, and unprocessed) in four normal-hearing listeners, with degree of simulated reverberation (no, mild, and severe reverberation) as a parameter. In the second experiment, vowel identification was measured as a function of the F0 difference between two simultaneous vowels in six normal-hearing listeners, with the number of vocoder channels (8 and 24 channels, and unprocessed) as a parameter.**

Results: **Reverberation was detrimental to F0 discrimination in conditions with fewer numbers of vocoder channels. Despite the reasonable F0DLs (<1 semitone) with 24- and 8-channel vocoder processing, listeners were unable to benefit from F0 differences between the competing vowels in the concurrent-vowel paradigm.**

Conclusions: **The overall detrimental effects of vocoder processing are probably are due to the poor spectral representation of the lower-order harmonics. The F0 information carried in the temporal envelope is weak, susceptible to reverberation, and may not suffice for source segregation. To the extent that vocoder processing simulates cochlear implant processing, users of current implant processing schemes are unlikely to benefit from F0 differences between competing talkers when listening to speech in complex environments. The results provide further incentive for finding a way to make the information from low-order, resolved harmonics available to cochlear implant users.**

(Ear & Hearing 2005;26;451–460)

Fundamental frequency (F0) information has long been thought to play an important role in perceptually segregating simultaneous and nonsimultaneous sources (for reviews see Bregman, 1990; Darwin & Carlyon, 1995). Studies of listeners with normal hearing have found that when a competing voice is present, listeners generally find it easier to understand the target voice if the competing voice has a different F0 (Assmann & Summerfield, 1990; 1994; Bird & Darwin, 1998; Brokx & Nooteboom, 1982; Darwin & Carlyon, 1995; de Cheveigné et al., 1997). Most models that use F0 differences to separate concurrent speech require explicit estimation of the F0 of either one or both sources (Assmann & Summerfield, 1990; Meddis & Hewitt, 1992). These models would predict that ambiguous F0 information leads to a deterioration in speech reception performance.

For normal-hearing listeners, the perception of voice pitch and the ability to discriminate different F0s are thought to rely primarily on temporal fine-structure information, in particular the information carried in peripherally resolved lower-order harmonics (e.g., Houtsma & Smurzynski, 1990; Plomp, 1967; Smith et al., 2002). Under normal circumstances, the frequencies of these harmonics are believed to be encoded by their place of excitation on the basilar membrane, by the temporal pattern of their auditory nerve responses, or by some combination of the two.

Cochlear implant users are unlikely to use the same F0 cues as normal-hearing listeners. The reasons for this are related to various properties of cochlear implants, which operate by bypassing the outer, middle, and inner ear to directly stimulate the auditory nerve. Most cochlear implant users today are implanted with multichannel devices (Clark et al., 1990; Loizou, 1999). 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 a small number of contiguous frequency bands or channels. The amplitude envelopes from each channel are low-pass filtered, typically at 400 Hz, and imposed on biphasic pulse carriers. The limited spectral resolution of current implant systems means that the lower harmonics of speech, which may give normal-hearing listeners spectral cues to pitch, are not represented individually and are therefore not resolved. Furthermore, the low-pass filtering of the envelopes eliminates most temporal fine-structure cues. However, voice pitch is in prin-

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

0196/0202/05/2605-0451/0 • Ear & Hearing • Copyright © 2005 by Lippincott Williams & Wilkins • Printed in the U.S.A.

ciple available in implant-processed speech via the periodicity in the temporal envelope (Green et al., 2002; Moore, 2003), so long as the cutoff frequency of the envelope-extraction filter is sufficiently high to pass the voice F0. McKay et al. (1994) showed that some implant users can detect differences, in the range of human voice pitch, as small as 2%, although most users have considerably higher difference limens. For instance, Geurts and Wouters (2001), using synthetic vowels to measure the F0 difference limens (F0DLs), found that CIS implant users could discriminate differences of between 4% and 13%. Such difference limens (see also Busby et al., 1993; Wilson, 1997) are an order of magnitude higher than those found in normal-hearing listeners when low-order (resolved) harmonics are present, but are only slightly higher than those found when only temporal envelope cues are presented to normal-hearing listeners (e.g., Bernstein & Oxenham, 2003; Burns & Viemeister, 1976; 1981; Kaernbach & Bering, 2001; Shackleton & Carlyon, 1994). Thus, normal-hearing listeners and cochlear implant users may share the same inability to efficiently code periodicity from information in the temporal envelope (Carlyon et al., 2002).

Another possible difficulty with envelope-periodicity F0 cues is their susceptibility to reverberation. The daily acoustic environments of implant users are often reverberant (e.g., living rooms, classrooms, music halls, and houses of worship). Previous studies examining the effects of cochlear implant processing on F0 discrimination were conducted under anechoic conditions (Faulkner et al., 2000; Fu et al., 1998b; Green et al., 2002), and the potential influence of reverberation has not been systematically examined. Reverberation has a "smearing" effect on envelope modulation (Houtgast et al., 1980; Steeneken & Houtgast, 1980) and is therefore likely to be detrimental to envelope-based F0 perception, even for steady-state sounds.

Cochlear implant users invariably exhibit poorer speech reception than normal. Although poor speech reception in implant users can be due to many factors, poor F0 information may be one reason why performance is particularly poor in complex, fluctuating backgrounds, in which listeners must perceptually segregate the target from the masker (Nelson et al., 2003; Nelson & Jin, 2004; Qin & Oxenham, 2003; Stickney et al., 2004). Specifically, poor F0 coding may result in the loss of F0 as a segregation cue. Although a link between F0 coding and speech segregation ability has been hypothesized before, no direct test has yet been undertaken.

The aim of this study was to address the question of a link between F0 coding and source segregation by examining the effects of using primarily envelope periodicity cues on F0 discrimination in anechoic and reverberant conditions (experiment 1) and on the ability to use F0 differences in segregating and identifying competing sound sources (experiment 2). We tested normal-hearing listeners using noise-excited envelope vocoder processing, as used in many previous studies (e.g., Dorman et al., 1997; Fu et al., 1998a; Rosen et al., 1999; Shannon et al., 1995; Shannon et al., 1998). This technique simulates certain aspects of cochlear implant processing, such as the loss of frequency selectivity, by filtering the stimulus into a small number of broadly tuned frequency channels, and the loss of temporal fine-structure information by using only the temporal envelope in each frequency channel to modulate noise carriers. Although such simulations clearly do not capture all aspects of cochlear implant perception (such as the limited dynamic range), they have certain advantages in that they avoid the large intersubject variability in the performance of actual implant users and that they can also be used to probe certain aspects of normal hearing.

F0 Difference Limens

The focus of this experiment was to examine the effects of noise-excited vocoder processing and reverberation on the ability of listeners to discriminate small changes in F0 between two sequentially presented harmonic tone complexes.

METHODS

Participants

Four normal-hearing listeners participated in this experiment (audiometric thresholds between 125 and 8000 Hz were <20 dB HL). They were undergraduate and graduate students with ages ranging from 19 to 28 yr.

Stimuli

All stimuli were digitally generated, treated, processed, and stored on computer disk using Matlab (Mathworks, Natick MA). The original stimuli were harmonic tone complexes, composed of equal-amplitude harmonics between 80 and 6000 Hz. The stimuli were first treated to simulate various reverberation conditions, and then processed to simulate the effects of cochlear implant sound processing.

Three conditions were tested. The first condition used sine-phase tone complexes, which simulate a pulsatile source, such as the human vocal folds, in an anechoic environment. The second condition convolved the sine-phase tone complexes with the recorded impulse response of a classroom (RT_{60} = 0.5 sec) to simulate the effects of mild reverberation.

The third condition used random-phase harmonic tone complexes, which were designed to simulate the phase relationships present in a highly reverberant environment. In using random-phase harmonic tone complexes we sought to simulate the worst-case scenario, that is, the most reverberant environment. Although randomizing the phase does not simulate all aspects of reverberation, it does result in greatly reduced average temporal envelope modulation, which is likely to be the most important cue for F0 discrimination based on temporal envelope properties.

Noise-excited envelope vocoder processing was used to simulate the effects of cochlear implant sound processing. The unprocessed stimuli were first bandpass filtered (6th-order Butterworth filters) into 1, 4, 8, 24, or 40 contiguous frequency channels between 80 and 6000 Hz. The entire frequency range was divided equally in terms of the Cam scale* (Glasberg & Moore, 1990). For instance, in the 24-channel condition, the filter with the lowest center frequency had lower and upper cutoff frequencies of 80 and 121.18 Hz, respectively. The bandwidths of the filters in the 24-channel condition are 1.16 Cams, which is only somewhat wider than the estimated bandwidths of human auditory filters (1 Cam, by the definition of Glasberg & Moore, 1990). To avoid differences in group delay between filters, zero-phase digital filtering was performed.† The envelopes of the signals were extracted by half-wave rectification and low-pass 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 F0 cues in the envelope as far as possible. The envelopes were then used to amplitude modulate independent white-noise carriers. The same bandpass filters that were used to filter the original stimuli were then used to filter the amplitude-modulated noises. Finally, the modulated narrowband noises were summed and scaled to have the same level as the unprocessed stimuli.

On each trial, the listener was presented with two successive stimulus tokens, separated by 200-msec pauses. Each stimulus token had a total duration of 200 msec and was gated on and off with 50-msec raised-cosine ramps. In the mild-reverberation condition, the stimuli were gated after reverberation had been added, so the total duration remained 200 msec. During each trial, one of the intervals contained the stimulus token with the nominal F0 ($F0_{ref}$), whereas the other interval contained the stimulus token with the comparison F0 ($F0_{ref}$ + $\Delta F0$). The order of presentation of the two intervals was selected randomly with equal probability from trial to trial. The $F0_{ref}$ was roved by $\pm 10\%$ from trial to trial to encourage listeners to compare the F0 of the two stimuli presented within each trial, rather than relying on an internal reference. Two nominal $F0_{ref}$ of 130 and 220 Hz were tested. These values were selected as they represent the mean F0 of male and female speech respectively. The stimulus levels were roved by ± 3 dB from interval-to-interval, around the mean overall level of 70 dB SPL, to minimize the effects of intensity on listener judgments.

Procedure

The F0DLs were measured using a two-alternative forced-choice paradigm. The one-up two-down adaptive procedure was used to track the 70.7% correct point (Levitt, 1971). At the beginning of a run, $\Delta F0$ was set to 20%. The value of $\Delta F0$ was reduced after two consecutive correct responses and increased after an incorrect response. The factor of variation of $\Delta F0$ was initially 1.58. It was reduced to 1.25 after the first reversal, and then to 1.12 after the next two reversals. Thresholds were calculated as the geometric mean of the $\Delta F0$ values at the last six reversals. A threshold measurement was considered out of range if a listener was repeatedly (in at least 3 of the 5 runs) unable to identify the higher-F0 interval at $\Delta F0$ values of 50%.

The experiment was conducted with the participant seated individually inside a double-walled soundproof booth. The preprocessed stimuli were played out via a sound card (LynxStudio LynxOne) with 24-bit resolution at a sampling frequency of 22.05 kHz. The stimuli were then passed through a programmable attenuator (TDT PA4) and headphone buffer (TDT HB6) before being presented diotically via a pair of Sennheiser HD580 headphones. The listeners were instructed to indicate the interval that contained the stimulus with the higher pitch. The two intervals were marked visually, and visual feedback was provided after each trial. The response on each trial was collected via a computer keyboard inside the sound booth.

All participants went through a training session, of approximately 2 hr, to familiarize them with the stimuli and experimental tasks. Each participant took part in five experimental sessions of approxi-

* This is more frequently referred to as the ERB scale. However, as pointed out by Hartmann (1997), ERB simply refers to equivalent rectangular bandwidth, which is not unique to 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.

† Zero-phase forward and reverse digital filtering was implemented with the use of the Matlab "filtfilt"command.

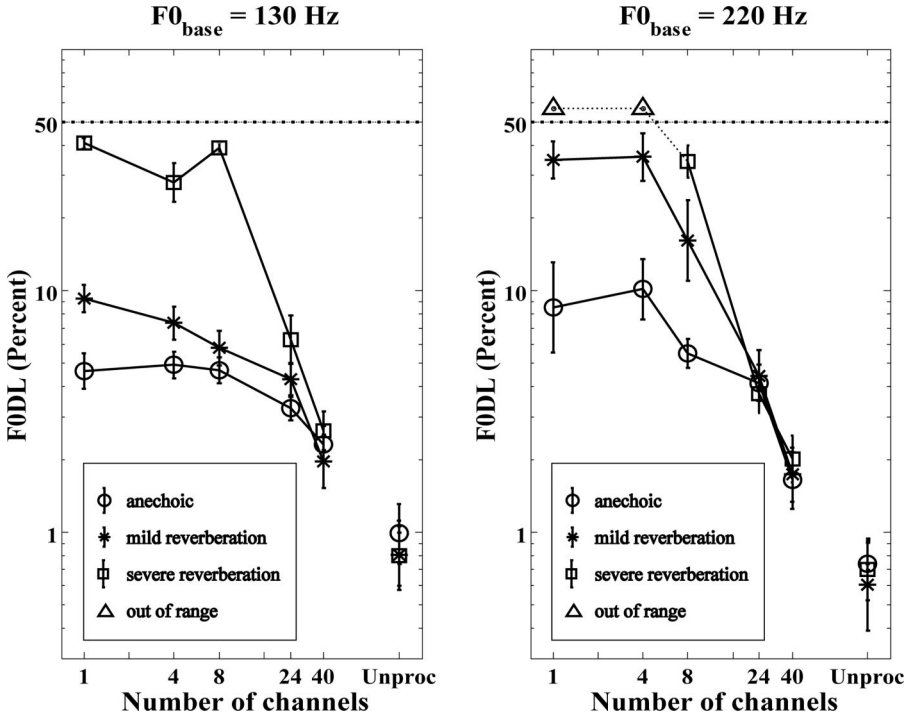

Fig. 1. Mean results from experiment 1. Each data point represents the mean F0DL (%) across four subjects; error bars denote ±1 SEM. Results from the different $F0_{ref}$ are shown on separate plots, $F0_{ref}$ = 130 Hz on the left and $F0_{ref}$ = 220 Hz on the right. Different reverberation conditions (i.e., no reverberation, mild reverberation, and severe reverberation) are shown as different symbols. Long-dashed horizontal lines show the measurement limit. Up-pointing triangles represent F0DLs outside the range of measurement.

mately 2 hr each. The five sessions took place over the span of 2 to 3 wk, depending on the availability of the participants. Each experimental session consisted of 36 runs, measuring the F0DL for each experimental condition (2 F0 conditions × 3 reverberation conditions × 6 vocoder processing conditions). Conditions were presented in a random order that varied across subjects and across repetitions. Individual thresholds were calculated as the geometric mean from the five repetitions of each condition.

RESULTS

The patterns of results from the individual subjects were very similar and so only the mean data are shown. Figure 1 shows the estimated F0DLs (expressed as a percentage of the $F0_{ref}$) as a function of the number of vocoder channels. Each point represents the geometric mean across subjects and the error bars denote ±1 SEM on the logarithmic scale. The results from $F0_{ref}$ = 130 Hz and $F0_{ref}$ = 220 Hz are shown in the left and right panels, respectively. The different reverberation conditions (i.e., nonreverberant, mild reverberation, and random phase) are shown as different symbols. The long-dashed line shows the measurement limit. The up-pointing triangles represent F0DLs outside the range of measurement.‡ Overall, F0DLs decreased with increasing number of vocoder channels. Furthermore, the detrimental effects of reverberation on F0DLs increased with decreasing number of channels.

A three-factor (F0, reverberation, and processing condition) within-subject analysis of variance (ANOVA) was performed on the log-transformed data. The following main effects and interactions were found to be significant. There was a significant main effect of processing condition [$F_{5,15} = 116.725$, $p < 0.001$], confirming the clear deterioration in performance with decreasing number of channels. Similarly, there was a main effect of reverberation [$F_{2,6} = 63.978$, $p < 0.001$]. Significant interactions were found between processing condition and reverberation [$F_{10,30} = 32.259$, $p < 0.001$], as well as between F0 and processing condition [$F_{5,15} = 18.186$, $p < 0.001$]. These interactions reflect the trends in the data for reverberation to be more detrimental at small channel numbers than at large (or unprocessed) and for thresholds to be poorer at 220 Hz than at 130 Hz, especially in the reverberant conditions with a small number of channels.

Our findings can be understood in the context of the different F0 cues available to the auditory system. When the spectral and temporal fine-structure cues in the processed stimuli are more representative of the original stimuli (i.e., with large numbers of channels) the F0 percept is probably driven by spectral and temporal fine-structure cues. When spectral and temporal fine-structure cues are weak (i.e., with small numbers of channels), the F0 percept is probably derived from temporal envelope cues. The pitch salience associated with temporal

‡ For the purposes of statistical analysis, out-of-range thresholds were set to the maximum allowable percentage (50%).

envelope cues is known to be weaker than that associated with spectral or fine-structure cues (e.g., Burns & Viemeister, 1976; 1981; Shackleton and Carlyon, 1994), and F0DLs are correspondingly larger. Furthermore, when listeners are forced to rely on envelope cues (i.e., with small numbers of channels) to perform F0 discrimination rather than fine-structure cues, the effects of reverberation are likely to be more detrimental. Reverberation can be characterized as a low-pass-filtering of the envelope modulation (e.g., Houtgast et al., 1980; Steeneken & Houtgast, 1980). Thus, reverberation is likely to smear out envelope-based F0 information, particularly at the higher (220 Hz) F0, where the envelope fluctuations are more rapid. The auditory system itself also seems to act as a low-pass filter in the modulation domain (e.g., Kohlrausch et al., 2000), which may explain why (with low channel numbers) thresholds were somewhat poorer in the 220-Hz condition than in the 130-Hz condition, even in the absence of reverberation.

Overall, the detrimental effects of vocoder processing may be attributed to the poor spectral and temporal fine-structure representation of the lower-order harmonics and to the disruption to envelope F0 cues caused by reverberation.

CONCURRENT-VOWEL IDENTIFICATION

As stated in the introduction, F0 information is believed to play an important role in perceptually segregating sound sources. Although, in principle, temporal envelope cues to voice pitch are available in implant processed speech (e.g., Busby et al., 1993; Geurts & Wouters, 2001; McKay et al., 1994; Wilson, 1997), less is known about the ability of listeners to use such cues for segregation. Scheffers (1983) showed that two vowels played simultaneously with different F0s were easier to understand than two vowels with the same F0. Though Scheffers' concurrent-vowel paradigm is not an accurate representation of the everyday situation, it does offer a well-controlled means of examining the contribution of F0 information to the segregation of two speech sounds. By presenting two synthetic vowels simultaneously at equal level, the effects of semantic, grammatical, and perceptual grouping cues, such as onset asynchronies, can be eliminated. Although the concurrent-vowel paradigm has been used to examine the utility of F0 information for source segregation with normal-hearing listeners (Assmann & Summerfield, 1990; 1994; Bird & Darwin, 1993; 1998; Brokx & Nooteboom, 1982; Culling & Darwin, 1993; Darwin & Carlyon, 1995; de Cheveigné et al., 1997; Summerfield and Assmann, 1991) and hearing-impaired listeners (Arehart et al., 1997), to our knowledge, the paradigm has not been used to examine the utility of F0 information with either actual or simulated cochlear implant users. This experiment examined the effects of vocoder processing on normal-hearing listeners' ability to use F0 information for source segregation. Vowel identification was measured as a function of the difference in F0 (ΔF0) between the two vowels (0, 1, 2, 4, 8, 12, and 14 semitones) and processing condition (unprocessed, 24-channel noise-excited envelope vocoder, and 8-channel noise-excited envelope vocoder).

TABLE 1. Formant frequencies (Hz) for vowels

	Male Formant			
Vowel	F1 (60)	F2 (90)	F3 (150)	F4 (200)
/i/	342	2322	3000	3657
/A/	768	1333	2522	3687
/u/	378	997	2343	3357
/ɛ/	580	1799	2605	3677
/ɝ/	474	1379	1710	3334

Values enclosed in parentheses represent formant bandwidths in Hz.

METHODS

Participants

Six native speakers of American English (audiometric thresholds between 125 and 8000 Hz were <20 dB HL) took part in this experiment. Their ages ranged from 19 to 22.

Stimuli

Five American English vowels (/i/ as in *heed*, /A/ as in *hod*, /u/ as in *hood*, /ɛ/ as in *head*, /ɝ/ as in *herd*) were synthesized by using an implementation of Klatt's cascade synthesizer (Klatt, 1980). They were generated at a sampling frequency of 20 kHz, with 16-bit quantization. The formant frequencies and bandwidths (see Table 1) used to synthesize the vowels were based on the estimates of Hillenbrand et al. (1995) for an average male talker. The vowels were chosen for their positions in the F1-F2 space and because their natural duration characteristics (House, 1960; 1961) are similar to the stimulus durations used in this experiment (200 msec). Each vowel was generated with seven different F0 (see Table 2). Changing the F0 maintaining the same

TABLE 2. Fundamental frequencies of the vowel pairs used in experiment 2, both in terms of absolute frequencies (in Hz) and frequency ratios (in semitones)

ΔF0 (semitones)	0	1	2	4	8	12	14
$F0_A$ (Hz)	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$F0_B$ (Hz)	100.0	105.9	112.2	126.0	158.7	200.0	224.5

formant frequencies is more akin to an intra-talker F0 difference than an inter-talker F0 difference; this was intended to eliminate potential confounding sources of speaker cues, such as other glottal source differences and vocal tract length (Bachorowski & Owren, 1999; Darwin et al., 2003).

The concurrent-vowel pairs were constructed by summing two single vowels with equal levels, with their onsets and offsets aligned, and with their pitch periods in phase at the onset of the stimulus. No vowel was paired with itself to generate the concurrent-vowel pairs. Each concurrent-vowel token was constructed by using one vowel with an F0 of 100 Hz and the other with an F0 of 100 Hz + ΔF0, where the ΔF0 ranged from 0 to 14 semitones (see Table 2). This yielded a total of 140 concurrent-vowel stimuli (10 vowel-pairs × 2 F0 combinations × 7 ΔF0s). Each stimulus had a total duration of 200 msec and was gated on and off with 25-msec raised-cosine ramps. The stimuli were presented at an overall level of 70 dB SPL.

All stimulus tokens were digitally generated, processed, and stored on computer disk before the experiments. The noise-excited vocoder processing used to simulate the effects of cochlear implant sound processing is the same as that used in the previous experiment. All vowels and concurrent vowel pairs were processed under 24-channel and 8-channel processing conditions.

Procedure

Both single-vowel and concurrent-vowel identification performance was measured by using a forced-choice paradigm. The listeners were instructed to identify the vowels heard by selecting visual icons associated with the vowels. In the single-vowel identification task, listeners were instructed to identify the vowel heard by selecting from five different choices. In the concurrent-vowel identification task, listeners were instructed to identify both the constituent vowels. Listener performance was scored as the percentage of correct responses. In the double-vowel identification task, a response was considered correct only if both vowels were correctly identified.

Each experimental session was broken into six blocks comprising the three vocoder-processing conditions with either the single vowels or concurrent vowels. Within each block, the presentation orders of the F0s (for single vowel identification) or F0 differences (for concurrent-vowel identification) were randomized. The single-vowel blocks contained 70 stimuli each (5 vowels × 7 F0 × 2 repetitions); the concurrent-vowel blocks contained 140 stimuli each (20 vowel-pairs × 7 ΔF0). During each trial, the responses were entered via a computer keyboard and mouse inside the booth. No feedback was provided.

Listeners were tested in a double-walled soundproof booth. The stimuli were played out via a sound card (LynxStudio LynxOne), passed through a programmable attenuator (TDT PA4) and headphone buffer (TDT HB6), before being presented diotically via a pair of Sennheiser HD580 headphones.

Every participant took part in two training sessions and three experimental sessions. The training sessions were designed to familiarize the participants with the experimental stimuli and the identification tasks. They were asked to perform the same task in the training sessions as in the experimental sessions. Before the experimental sessions, all participants were required to achieve at least 90% identification accuracy in the single-vowel task. The five sessions took place over the span of 1 to 2 wk, depending on the availability of the participants.

RESULTS

Figure 2 shows the identification accuracy as a function of F0 in the single-vowel identification task (dotted lines; shown as semitones above 100 Hz) and difference between the F0 of the constituent vowels in the concurrent-vowel identification task (solid lines). The unprocessed conditions are shown in the left panel, the 24-channel conditions in the center panel, and the 8-channel conditions in the right panel. For presentation purposes, the results are pooled across subjects, across vowels (or vowel-pairs) and across vowel order (i.e., which vowel had the higher F0). Some more fine-grained analysis is provided later.

To investigate trends in the concurrent-vowel data, a within-subject ANOVA with two factors (ΔF0 and processing condition) was conducted. All scores were arcsine transformed§ (Studebaker, 1985) before analysis. The ANOVA analysis showed significant main effects of ΔF0 [$F_{6,30} = 10.343$, $p < 0.001$] and processing condition [$F_{2,10} = 42.158$, $p < 0.001$] and an interaction between ΔF0 and processing condition [$F_{12,60} = 10.150$, $p < 0.001$]. Post hoc comparison, using Bonferroni correction, showed that the mean score difference between the 24-channel condition (58.7%) and the 8-channel condition (45.3%) was significant ($p < 0.05$), as was the difference between the unprocessed condition and the two processed conditions ($p < 0.05$ in both cases).

§ The rationale for arcsine transformation is that the data of percent correct have nonuniform variance, whereas the transformed data have the property of stabilized variance of binomial data and thus are more suitable for analysis of variance (ANOVA) and other statistical analyses.

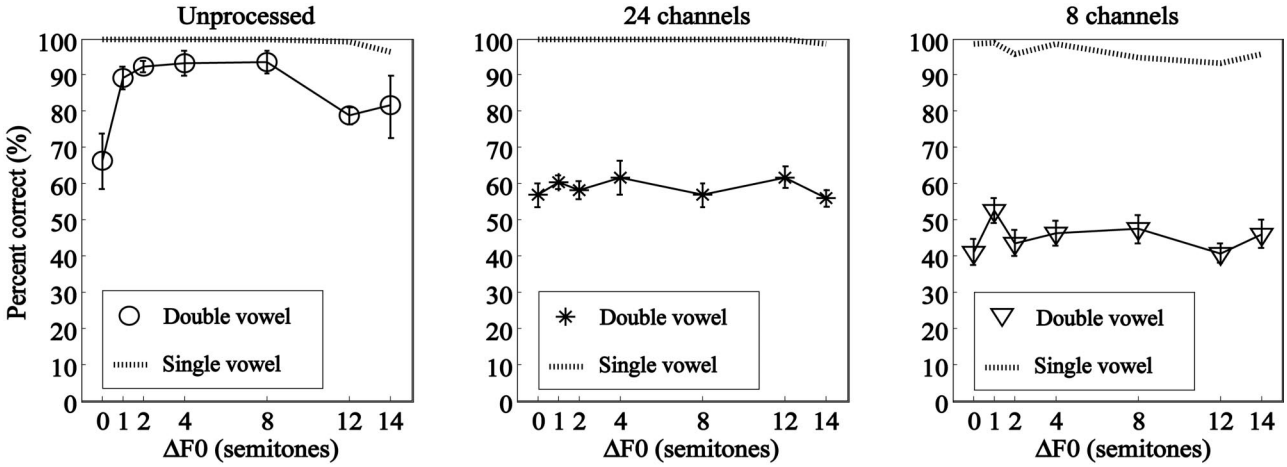

Fig. 2. Mean results from experiment 2. Dotted lines show the percent of correct responses as a function of the F0 in the single-vowel identification task, where F0 is described in terms of the number of semitones above 100 Hz. Solid lines show the percent of correct responses as a function of the ΔF0 (in semitones) between constituent vowels in the concurrent-vowel identification task, where the lower F0 was always 100 Hz. Error bars denote ±1 SEM. Unprocessed conditions are shown in the left panel, the 24-channel conditions in the center panel, and the 8-channel condition in the right panel.

In the unprocessed conditions (Fig. 2, left panel), listeners show an average 26 percentage-point improvement in performance as ΔF0 increases from 0 to 2 semitones, after which their performance plateaus until the ΔF0 equals 12 semitones (one octave), consistent with Brokx & Nooteboom (1982). As expected, when the ΔF0 equals one octave the harmonics of the two constituent vowels become inseparable, leading to a drop in identification performance. At ΔF0 of 14 semitones, the identification performance seems to improve somewhat, although the performance difference between 12 and 14 semitones was not statistically significant ($p > 0.05$).

In contrast to the unprocessed conditions, ΔF0 had no significant effect on performance in the concurrent-vowel identification task in either of the vocoder-processed conditions (separate ANOVA; $p > 0.05$). A possible explanation for the lack of ΔF0 effect in the vocoder-processed conditions may lie in the inability of the auditory system to use envelope cues to extract the F0s of two periodic stimuli that excite the same region of the cochlea. Carlyon (1996) and Carlyon et al. (2002) have shown that listeners fail to hear two underlying pitches when mixtures of two periodic pulse trains are either applied to a single implant electrode or presented to normal-hearing listeners after filtering out the lower harmonics. Our findings extend those of Carlyon and colleagues by showing that even with substantial spectral differences between the two sources (presumably leading to spectral formant regions where one or the other F0 dominates) listeners were not able to use these envelope-based F0 cues for segregation.

In the unprocessed conditions, the benefits of a difference in F0 were evident for all vowel pairs. However, levels of performance differed among vowel pairs. The largest ΔF0 benefit was seen when constituent vowels had similar first formant (F1) or second formant (F2) values (e.g., /herd, head/ and /herd, hod/). For example, /herd, head/ went from 38% at ΔF0 of 0 semitones to 85% at ΔF0 of 2 semitones; /herd, hod/ went from 41% at ΔF0 of 0 semitones to 75% at ΔF0 of 2 semitones. The least ΔF0 benefit was seen when constituent vowels have dissimilar F1 and F2 values (e.g., /heed, hod/). This probably is due to the already high performance (80%) at ΔF0 of 0 semitones.

With vocoder processing, although listeners had no trouble identifying any of the vowels presented in isolation (mean percent correct >90%), they experienced difficulties with concurrent vowel identification. In the concurrent-vowel identification task, whereas listeners on average performed above chance (>10%), they had a great deal of trouble when constituent vowels had similar F1 or F2 values. For example, in the 24-channel condition, the mean correct identification of the vowel pair /herd, head/ was ~40%, whereas /herd, hod/ was ~20%. In the 8-channel condition, the mean correct identification of the vowel pair /herd, head/ was ~30%, whereas /herd, hod/ was ~10%. When constituent vowels had dissimilar F1 and F2 values (e.g., /heed, hod/) the mean correct identification of the vowel pairs were well above the average for the condition.

DISCUSSION

In the first experiment, F0DLs were measured as a function of number of envelope-vocoder frequency

channels, with degree of simulated reverberation as a parameter. It was found that sensitivity to small F0 differences decreased with decreasing number of vocoder channels, and the detrimental effects of reverberation on F0DLs increased with decreasing number of channels (Fig. 1). In the second experiment, vowel identification was measured as a function of the F0 difference in a concurrent-vowel paradigm, with the number of vocoder channels as a parameter. Under vocoder processing, listeners were unable to benefit from F0 differences, even when the ΔF0 between concurrent vowels was as large as 8 semitones (Fig. 2). In contrast, under the same vocoder processing conditions, F0DLs of less than 1 semitone (or <6%) were found (Fig. 1).

A possible explanation for the apparent contrast between the findings of the two experiments may lie in the limits of the auditory system's ability to extract the F0 of two periodic stimuli presented simultaneously. In experiment 1, listeners were asked to discriminate F0 differences of sequentially presented tone complexes. In experiment 2, listeners were asked to identify simultaneously presented vowels. It is conceivable that, although the auditory system may be able to use envelope periodicity cues to extract the F0 of one periodic stimulus (as in experiment 1), it may be incapable of extracting the F0s of two simultaneously presented stimuli (as is required in experiment 2). Indeed, although the question of whether one or two sources were heard was not formally addressed in this study, unsolicited comments by several subjects suggested that if given the choice they would have often reported hearing only one vowel during the concurrent-vowel conditions. Informal listening by the authors also suggested that only one source was perceived in most of the vocoder-processed concurrent-vowel conditions.

In agreement with the present study, Deeks and Carlyon (2004) found no clear advantage for having their masker and target speech on different carrier rates, when the masker and target speech were passed through the same frequency channels. They went further to show that the different carrier rates were not a useful cue for segregation, even when the masker and target speech excited separate channels. If the ability of listeners to extract the F0 from sequentially and simultaneously presented stimuli are indeed different, then the F0DLs found with sequentially presented stimuli would not be an appropriate indicator of a listener's ability to use F0 cues for speech segregation.

An alternative explanation is that it is not the mode of presentation (sequential versus concurrent) but rather the nature of the stimuli (equal-amplitude sine-phase harmonics versus vowel-shaped harmonics with phase shifts dependent on the vowel filtering) that explains the different outcomes of experiments 1 and 2. However, informal examinations of the temporal envelopes after vocoder processing suggest that both types of stimuli give similarly well-defined temporal envelopes.

It is worth reiterating here that although studies using normal-hearing listeners and noise-excited envelope vocoders have their advantages, the inherent differences between acoustic and electrical stimulation (Litvak et al., 2001; Rubenstein et al., 1999; Throckmorton & Collins, 2002) mean that the results from simulation studies should be interpreted in terms of trends rather than quantitative estimates of implant user performance. Taken in this light, the current results nevertheless suggest severe limits in the perceptual use of the envelope F0 information for cochlear implant users. In the absence of other (auditory and nonauditory) cues, implant users are unlikely to benefit from differences in F0 between competing sources. Although our current experiments do not address the question of whether F0 can be used as an additional cue in conjunction with others, our results are in line with predictions from earlier studies that poor F0 representation may underlie some of the difficulties experienced by implant users in complex environments (e.g., Qin & Oxenham, 2003).

SUMMARY

1. Simulated cochlear implant processing, using a noise-excited vocoder, had a detrimental effect on listeners' F0 discrimination abilities. Under processed conditions, performance worsened with decreasing numbers of vocoder frequency channels.

2. Reverberation did not affect F0 discrimination in unprocessed conditions, but was detrimental to F0 discrimination in processed conditions with fewer numbers of vocoder channels. The effect of reverberation was particularly marked at the higher $F0_{ref}$ (220 Hz), in line with expectations based on temporal-envelope processing.

3. Despite the reasonable F0DLs (<1 semitone) with 24- and 8-channel vocoder processing in a sequential paradigm, listeners were unable to benefit from F0 differences between the competing vowels in a concurrent-vowel paradigm.

4. The overall detrimental effects of vocoder processing are probably due to the poor representation of the lower-order harmonics. The present study provides further incentive for finding ways of making the information from low-order, resolved harmonics available to cochlear implant users.

5. To the extent that vocoder processing simulates cochlear implant processing, results from this and

other studies provide no evidence that users of current implant processing schemes can benefit from F0 differences between competing talkers. However, it remains possible that F0 information encoded in the envelope fluctuations may provide additional help in real-world situations, where visual, linguistic, and other auditory cues are also present.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (NIDCD grant R01 DC 05216). Barbara Shinn-Cunningham kindly provided the room impulse response used in experiment 1. We thank Christophe Micheyl, Andrea Simonson, Joshua Bernstein, and Evan Chen for helpful comments on an earlier version of this manuscript. We also thank Rosalie Uchanski, John Culling, Bob Carlyon, and an anonymous reviewer for their many helpful suggestions.

Address for correspondence: Michael K. Qin, Naval Submarine Medical Research Laboratory, Box 900, Groton, CT 06349. E-mail: qin@nsmrl.navy.mil.

Received June 21, 2004; accepted March 30, 2005.

REFERENCES

Arehart, K. H., King, C. A., & McLean-Mudgett, K. S. (1997). Role of fundamental frequency differences in the perceptual separation of competing vowel sounds by listeners with normal hearing and listeners with hearing loss. *Journal of Speech, Language, and Hearing Research*, *40*, 1434–1444.

Assmann, P. F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, *88*, 680–697.

Assmann, P. F., & Summerfield, Q. (1994). The contribution of waveform interactions to the perception of concurrent vowels," J. *Journal of the Acoustical Society of America*, *95*, 471–484.

Bachorowski, J. A., & Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *Journal of the Acoustical Society of America*, *106*, 1054–1063.

Bernstein, J. G., & Oxenham, A. J. (2003). Pitch discrimination of diotic and dichotic tone complexes: harmonic resolvability or harmonic number? *Journal of the Acoustical Society of America*, *113*, 3323–3334.

Bird, J., & Darwin, C. J. (1998). Effects of a difference in fundamental frequency in separating two sentences. In A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis (Eds.) *Psychophysical and Physiological Advances in Hearing.* London: Whurr.

Bregman, A. S. (1990). Auditory Scene Analysis: The Perceptual Organisation of Sound. Cambridge, MA: Bradford Books, MIT Press.

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Phonetics*, *10*, 23–36.

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.

Carlyon, R. P. (1996). Encoding the fundamental frequency of a complex tone in the presence of a spectrally overlapping masker. *Journal of the Acoustical Society of America*, *99*, 517–524.

Carlyon, R. P., van Wieringen, A., Long, C. J., Deeks, J. M., & Wouters, J. (2002). Temporal pitch mechanisms in acoustic and electric hearing. *Journal of the Acoustical Society of America*, *112*, 621–633.

Clark, G. M., Tong, Y. C., & Patrick, J. F. (1990). *Cochlear Prostheses*. Churchill Livingston, Edinburgh, 1990.

Culling, J. F., & Darwin, C. J. (1993). Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0. *Journal of the Acoustical Society of America*, *93*, 3454–3467.

Culling, J. F., & Darwin, C. J. (1994). Perceptual and computational separation of simultaneous vowels: Cues arising from low-frequency beating. *Journal of the Acoustical Society of America*, *95*, 1559–1569.

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *Journal of the Acoustical Society of America*, *114*, 2913–2922.

Darwin,C. J., & Carlyon,R. P. (1995). Auditory Grouping. In B. C. J. Moore (Ed.) Hearing. San Diego, CA: Academic Press.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., & Aikawa, K. (1997). Concurrent vowel identification. I. *Effects of relative amplitude and Fo difference.*, Journal of the Acoustical Society of America, *101*, 2839–2847.

Deeks, J. M., & Carlyon, R. P. (2004). Simulations of cochlear implant hearing using filtered harmonic complexes: implications for concurrent sound segregation. *Journal of the Acoustical Society of America*, *115*, 1736–1746.

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.

Fu, Q. J., Shannon, R. V., & Wang, X. S. (1998a). 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.

Fu, Q. J., Zeng, F. G., Shannon, R. V., & Soli, S. D. (1998b). 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.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*, 103–138.

Green, T., Faulkner, A., & Rosen, S. (2002). Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling cochlear implants. *Journal of the Acoustical Society of America*, *112*, 2155–2164.

Hartmann,W. M. (1997). Signals, Sound, and Sensation. New York, NY: Springer-Verlag.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099–3111.

House, A. S. (1960). Formant band widths and vowel preference. *Journal of Speech and Hearing Research*, *3*, 3–8.

House, A. S. (1961). On vowel duration in English. *Journal of the Acoustical Society of America*, *33*, 1174–1178.

Houtgast, T., Steeneken, H. J. M., & 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., & Smurzynski, J. (1990). Pitch identification and discrimination for complex tones with many harmonics. *Journal of the Acoustical Society of America*, *87*, 304–310.

Kaernbach, C., & Bering, C. (2001). Exploring the temporal mechanism involved in the pitch of unresolved harmonics. *Journal of the Acoustical Society of America*, *110*, 1039–1048.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, *67*, 971–995.

Kohlrausch, A., Fassel, R., & Dau, T. (2000). The influence of carrier level and frequency on modulation and beat- detection thresholds for sinusoidal carriers. *Journal of the Acoustical Society of America*, *108*, 723–734.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, *49*, 467–477.

Litvak, L., Delgutte, B., & Eddington, D. (2001). Auditory nerve fiber responses to electrical stimulation: Modulated and unmodulated pulse trains. *Journal of the Acoustical Society of America*, *110*, 368–379.

Loizou, P. C. (1999). Introduction to cochlear implants. *IEEE Engineering in Medicine and Biology Magazine*, *18*, 32–42.

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

Meddis, R., & Hewitt, M. (1992). Modeling the identification of concurrent vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, *91*, 233–245.

Moore, B. C. (2003). Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants. *Otology and Neurotology*, *24*, 243–254.

Nelson, P. B., & Jin, S. H. (2004). Factors affecting speech understanding in gated interference: Cochlear implant users and normal-hearing listeners. *Journal of the Acoustical Society of America*, *115*, 2286–2294.

Nelson, P. B., Jin, S. H., Carney, A. E., & Nelson, D. A. (2003). Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners. *Journal of the Acoustical Society of America*, *113*, 961–968.

Plomp, R. (1967). Pitch of complex tones. *Journal of the Acoustical Society of America*, *41*, 1526–1533.

Qin, M. K., & Oxenham, A. J. (2003). Effects of simulated cochlear implant processing on speech reception in fluctuating maskers. *Journal of the Acoustical Society of America*, *114*, 446–454.

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.

Rubenstein, J. T., Wilson, B. S., Finley, C. C, & Abbas, P. J. (1999). Pseudospontaneous activity: Stochastic independence of auditory nerve fibers with electrical stimulation. *Hearing Research*, *127*, 108–118.

Scheffers,M. T. M. (1983). Sifting vowels: Auditory pitch analysis and sound segregation, Ph.D. Thesis, Groningen University, The Netherlands.

Shackleton, T. M., & Carlyon, R. P. (1994). The role of resolved and unresolved harmonics in pitch perception and frequency-modulation discrimination. *Journal of the Acoustical Society of America*, *95*, 3529–3540.

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.

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

Steeneken, H. J. M., & Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *Journal of the Acoustical Society of America*, *69*, 318–326.

Stickney, G., Zeng, F.-G., Litovsky, R., & Assmann, P. (2004). Cochlear implant speech recognition with speech maskers. *Journal of the Acoustical Society of America*, *116*, 1081–1091.

Studebaker, G. (1985). A 'rationalized' arcsine transform. *Journal of Speech and Hearing Research*, *28*, 455–462.

Summerfield, A. Q., & Assmann, P. F. (1991). Perception of concurrent vowels: effects of pitch-pulse asynchrony and harmonic misalignment. *Journal of the Acoustical Society of America*, *89*, 1364–1377.

Throckmorton, C. S., & Collins, L. M. (2002). The effect of channel interactions on speech recognition in cochlear implant subjects: predictions from an acoustic model. *Journal of the Acoustical Society of America*, *112*, 285–296.

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

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