

Research paper

Acoustic model investigation of a multiple carrier frequency algorithm for encoding fine frequency structure: Implications for cochlear implants [☆]

Chandra S. Throckmorton, M. Selin Kucukoglu, Jeremiah J. Remus, Leslie M. Collins *

Department of Electrical and Computer Engineering, Duke University, 130 Hudson Hall, P.O. Box 90291, Durham, NC 27708-0291, United States

Received 24 January 2006; received in revised form 24 March 2006; accepted 27 March 2006

Available online 23 June 2006

Abstract

Current cochlear implants provide frequency resolution through the number of channels. Improving resolution by increasing channels is limited by factors such as the physiological feasibility of increasing the number of electrodes, the inability to increase the number of channels for those already implanted, and the increased possibility of channel interactions reducing channel efficacy. Recent studies have suggested an alternative method: providing a continuum of pitch percepts for each channel based on the frequency content of that channel. This study seeks to determine the frequency resolution necessary for the highest performance gain, which may give some indication of the feasibility for implementation in implants. A discrete set of carrier frequencies, instead of a continuum, are evaluated using an acoustic model to measure speech recognition. Performance increased as the number of available frequencies increased, and substantive improvement was seen with as few as two frequencies per channel. The effect of variable frequency discrimination was also assessed, and the results suggest that frequency modulation can still provide benefits with poor frequency discrimination on some channels. These results suggest that if two or more discriminable frequencies per channel can be generated for cochlear implant subjects then an improvement in speech recognition may be possible.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Frequency discrimination; Frequency modulation; Acoustic model; Cochlear implant

1. Introduction

That cochlear implants restore some level of hearing to most deaf individuals is well established. Recent advances in device hardware, surgical techniques, and speech processing strategies have resulted in excellent speech recognition for many recipients. However, speech recognition abilities still vary widely across subjects and the mechanisms responsible for this variability are only partially understood. Factors such as electrode design, electrode placement, analog versus pulsatile stimulation, peripheral nerve survival, central auditory system integrity, speech-processing strategy, and complex and/or unexpected current paths within or around the target neural tissue may play a role in this variability. Unfortunately, these factors are complex and difficult to control or assess in implanted subjects.

Abbreviations: CIS, continuous interleaved sampling; FAME, frequency amplitude modulation encoding; F1, F2, Formant frequencies (F1 indicates first formant, etc.); MCFA, multiple carrier frequency algorithm; SNR, signal to noise ratio; STFT, short-time fourier transform

[☆] Portions of this work were presented in “Encoding fine structure for improved speech understanding in cochlear implant subjects”, Association for Research in Otolaryngology Midwinter Meeting, New Orleans, LA, Feb. 2005. Also in “Encoding additional frequency information via variable pulse rate for improved speech understanding in cochlear implant subjects”, Conference on Implantable Auditory Prostheses, Asilomar, CA, August 2005

* Corresponding author. Tel.: +1 919 660 5260; fax: +1 919 660 5293.

E-mail addresses: cst@ee.duke.edu (C.S. Throckmorton), msk5@ee.duke.edu (M. Selin Kucukoglu), jjr6@ee.duke.edu (J.J. Remus), lcollins@ee.duke.edu (L.M. Collins).

Studies utilizing acoustic models provide a method by which trends in speech recognition performance can be measured while still retaining control over the underlying variables. Further, the use of normal-hearing subjects avoids demographic complications that might influence results from cochlear implant subjects (e.g. Shipp and Nedzelski, 1995; Geier et al., 1999; Rubinstein et al., 1999; van Dijk et al., 1999). However, considering the results from acoustic models in terms of predicting the performance of implant subjects must be approached with caution. The neural responses of electric and acoustic stimuli inherently differ (e.g. Kiang and Moxon, 1972; Hartmann et al., 1984). Thus, acoustic model results are useful in terms of trends in predicted performance rather than as exact predictions of the magnitude of performance changes.

Although Shannon et al. (1995) found that normal hearing subjects could achieve high speech recognition scores in quiet with only four channels, the results from several other acoustic model studies testing speech recognition in quiet have suggested that increasing the frequency information by increasing the number of channels improves performance (Dorman et al., 2000; Dorman et al., 1997; Dorman et al., 2002a,b). However, maximum speech recognition performance was achieved with a relatively small number of channels, typically fewer than 12.

In less than ideal conditions, the increase in the number of channels required to achieve maximum performance in speech recognition has also been noted, e.g. speech in noise (Dorman et al., 1998; Friesen et al., 2001; Fu et al., 1998), poor intensity resolution (Loizou et al., 2000), and degraded spectral information (Baskent and Shannon, 2003; Fu and Shannon, 1999). However, increasing the number of channels may not be realistic for subjects who have already been implanted, and if increasing the number of channels leads to increased channel interactions, speech recognition may not be improved. Zwolan et al. (1997) demonstrated that removing channels can actually be beneficial for speech recognition, in this case, as a method to compensate for indiscriminable electrodes.

Given the limitations on increasing the number of channels, recent research has focused on evaluating the performance of alternative methods for encoding additional frequency information. Some speech processing algorithms initially relied on the extraction of formant frequencies which were then encoded via electrode place and stimulation rate (Blamey et al., 1987; Clark, 1987; Patrick and Clark, 1991); however, these algorithms are no longer utilized due to their tendency to make errors in formant estimation in noisy environments, their inability to adequately represent non-speech stimuli such as music, and their poorer performance compared to more recent algorithms that convey the amplitude modulation of the spectrum (e.g. McKay et al., 1992).

Recently, several studies utilizing acoustic models have suggested that if each individual channel could provide additional frequency resolution, then speech recognition might be improved without increasing the number of chan-

nels (Nie et al., 2005; Zeng et al., 2005). Nie et al. (2005) proposed the Frequency Amplitude Modulation Encoding (FAME) algorithm which modulates the frequency by extracting the instantaneous frequency within each channel and adapting the carrier frequency based on that information. Nie et al. (2005) report that utilization of the FAME strategy can improve speech recognition by as much as 71% in normal hearing subjects listening through an acoustic model in noise. Using a similar technique, Zeng et al. (2005) demonstrated that increasing frequency information through the number of channels was not always equivalent to increasing frequency information within channel. For a Mandarin tone recognition task, a 32-channel amplitude modulation model (frequency information coded solely by number of channels) did not reach the performance level of an 8-channel model that included frequency modulation within channel. For the other tasks (speech recognition in quiet, speech recognition in noise, and speaker identification), at least 16 channels were required for the model presenting only amplitude modulation to match the performance level of the 8-channel model that modulated frequencies within channel. The implications of these studies are that if a method could be developed for increasing frequency resolution within channel for cochlear implant subjects, then speech recognition might be improved. Both studies propose using pulse rate to achieve this goal in implant subjects. The hypothesis is that each electrode provides a base pitch, and pulse rate can be adjusted to increase or decrease that particular pitch, thus providing more pitch percepts per channel. In acoustic models, this is limited to choosing frequencies within each channel's frequency range to represent the different possible pitches.

Although these results have only been demonstrated to date with acoustic models, Fearn (2001) also proposed a variable pulse rate speech processing strategy that was tested in implant subjects. This approach differs from FAME in that it only uses variable pulse rates for channels coding information below 1000 Hz. It also differs from FAME in the methods used to extract the frequency and code the frequency as pulse rate. This approach has been evaluated in a small group of implant subjects for which both speech and music perception were evaluated. Initial results indicated that speech recognition was not improved using this algorithm, but that music perception did improve, suggesting that using pulse rates to increase frequency resolution within channels may be possible.

Thus, results reported to date indicate that utilization of variable pulse rates show some promise as a method for encoding frequency information and that the additional information provided by a continuous frequency modulation of the presentation frequency appears to aid speech recognition. In this study, an alternative strategy is investigated that does not utilize the continuum of frequencies that is available in each channel under the FAME model. The proposed strategy associates a total of N_r discrete carrier frequencies with each analysis band that are then modulated by the temporal envelope. In any given analysis

window, only one presentation frequency is utilized, but this frequency can vary across analysis windows. The parameter N_r is varied systematically in order to ascertain relative performance gains associated with each increase in complexity. This simpler model may result in an easier implementation in an actual cochlear implant system since there would be a small number of predefined pulse rates to encode fine frequency structure. This study is performed in normal-hearing subjects listening to speech processed by a standard acoustic model. For ease in notation, this speech processing technique is termed multi-carrier frequency algorithm, or MCFA.

In addition to considering whether speech recognition performance could be improved using MCFA and assessing the relative complexity required to obtain a given level of performance improvement, a second study was performed to investigate the relative effect of variable psychophysical characteristics. Specifically, variable frequency discrimination ability was assessed as a function of channel. As originally posed, both FAME and MCFA essentially assume infinite frequency discrimination capabilities, both in the range of frequencies that can be discriminated and in the number of discriminable frequency steps that are available. The second study assesses the impact of limiting the discriminability of frequency on a channel-by-channel basis as well as whether the MCFA approach can still provide benefit when such frequency discrimination limitations are present.

2. Acoustic models

In this section, the implementation details associated with the acoustic models used to perform the studies are described. All models follow the same general design as the model described in Section 2.1; thus, only modifications to the design are described in the subsequent subsections. Each of the models were derived from the CIS-like model described in detail by Throckmorton and Collins (2002), and were implemented in MATLAB[®].

2.1. Standard model

Prior to processing by a particular acoustic model, speech tokens were pre-filtered for spectral equalization using a first-order high-pass Butterworth filter with cutoff frequency of 1 kHz. Anti-aliasing filtering was then performed by filtering the signals with a low-pass sixth-order Butterworth filter. The algorithm used to process the speech was designed to be similar to that performed by the CIS strategy (Kessler, 1999; Wilson et al., 1991). The pre-filtered speech was passed through a bank of eight or sixteen sixth order Chebyshev type I bandpass filters with central frequencies logarithmically spaced between 150 and 6450 Hz. The 8-channel model was considered the standard from which the other models were designed, while the 16-channel model was used in Experiment 2 to evaluate the difference between increasing frequency information

through increasing channel and increasing information through available presentation frequencies.

Envelopes of the bandpass filters were extracted using full-wave rectification and low-pass filtering, with this stage of filtering performed using eighth-order Chebyshev type I filters. Acoustic signals were generated by summing signals associated with all eight or sixteen bandpass filters. Carrier sinusoids were used whose frequencies were associated with the cutoff frequencies of the band (see Throckmorton and Collins, 2002), and amplitudes were set based on the root-mean-square energy of the envelope. Speech was analyzed in 2 ms analysis windows, similar to the method described in Dorman et al. (1997). The 8- and 16-channel models provided baseline results, as discussed in the Methods section.

2.2. FAME

An 8-channel implementation of FAME was developed that was based on the description provided in Nie et al. (2005). The envelope was extracted using the same method as the standard model, and frequency extraction was performed using Flanagan's vocoder approach (Flanagan, 1980; Flanagan and Golden, 1966). In the original FAME implementation, the carrier signal associated with each band was a sinusoid whose amplitude and frequency could be varied at each sampling instant based on the extracted envelope and frequency (see Eq. (1) in Nie et al. (2005)). In order to be more consistent with the other acoustic models, which used a 2 ms sampling window, the version of FAME that was implemented for this study utilized the same envelope extraction and analysis window approach as defined for the standard model. However, the carrier frequency that was used in each channel was determined using the approach defined in Nie et al. (2005), and this carrier frequency was used throughout the entire 2 ms analysis window.

2.3. MCFA

To implement MCFA, the 8-channel standard model was modified in the following way (see Fig. 1). The most prominent frequency in each analysis band was estimated for each 2 ms analysis window using the short-time Fourier transform (STFT). These frequencies were then mapped to the single, closest pre-defined presentation frequency. Each channel had an option of $N_r = 1, 2, 4,$ or 8 pre-defined presentation frequencies, depending on the MCFA model under consideration. The frequencies chosen for each channel were center frequencies of linear divisions of the passbands (see Table 1). Frequencies determined by the STFT to fall within a linear division were then mapped to the corresponding center frequency. Although, linear mapping was utilized for simplicity, other mappings might be equally (or more) appropriate. However, the goal of this work was to assess the effect of increasing frequency information rather than to assess the effect of coding

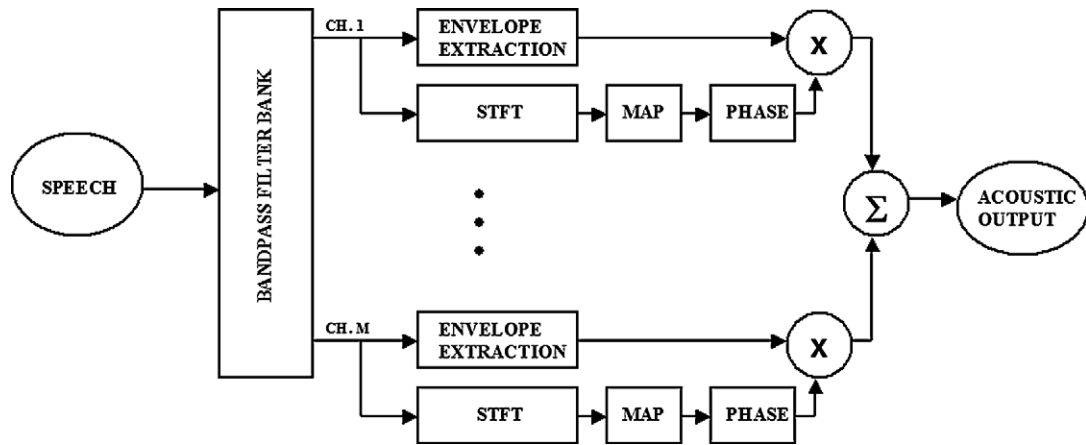


Fig. 1. A block diagram of an M-channel MCFA algorithm. In this study, $M = 8$ for all MCFA- n algorithms. Speech passes through a bandpass filter bank to separate information by channel. In order to determine the carrier frequency, a short-time Fourier transform (STFT) is performed, and the resulting frequency estimate is mapped to the nearest presentation frequency in the frequency set for that channel. The phase of the carrier is then adjusted to minimize amplitude variation between the current and previous analysis window. The carriers are amplitude modulated with the envelope information (extracted through full-wave rectification and a low-pass filter) and then summed to form the final acoustic output.

Table 1

The carrier frequencies (in Hz) to which the STFT frequency estimates were mapped for each MCFA- n model

Models	Ch. 1	Ch. 2	Ch. 3	Ch. 4	Ch. 5	Ch. 6	Ch. 7	Ch. 8
MCFA-2	173	276	442	707	1132	1810	2897	4636
	218	348	557	892	1427	2283	3653	5845
MCFA-4	161	258	413	661	1058	1692	2708	4333
	184	294	471	753	1205	1928	3086	4938
	206	330	528	846	1353	2165	3464	5543
	229	366	586	938	1500	2401	3842	6148
MCFA-8	156	249	398	638	1021	1633	2614	4182
	167	267	427	684	1095	1751	2803	4485
	178	285	456	730	1168	1869	2992	4787
	189	303	485	776	1242	1987	3181	5089
	201	321	514	823	1316	2106	3370	5392
	212	339	543	869	1390	2224	3559	5694
	223	357	572	915	1463	2342	3748	5996
	234	375	601	961	1537	2460	3937	6299

specific frequencies. Since implementation in cochlear implant recipients is unlikely to result in the exact mapping of pulse rate to specific frequencies, the choice of the frequencies was considered to be less important to the results than the number of available frequencies.

In the terminology used below, MCFA- n refers to the MCFA algorithm using $N_r = n$ potential carrier frequencies in each analysis band. In the case of only one pre-defined frequency per channel, the MCFA model was identical to the standard model. The choice of presentation frequency could change from analysis window to analysis window, but only one presentation frequency was used in each window. Changing frequency between analysis windows can lead to abrupt (i.e. noisy) amplitude transitions due to phase mismatch at the window boundaries; therefore, after the selection of carrier frequency, a phase adjustment was made to minimize abrupt amplitude variation. This is a concern specific to the use of acoustic models

and would likely not need to be addressed in the implementation for implant subjects.

The envelopes were extracted and the carrier signals were amplitude modulated using the same procedure as the standard model. The results of the processing associated with each analysis band were added together to form the resultant acoustic signal.

2.4. Impaired MCFA

This model simulates the impact of having two of the eight channels unable to fully exploit the MCFA processing. To simulate an extreme test case, it was assumed that only one frequency could be discriminated in two of the channels, while multiple frequencies were available in the remaining channels. Thus, two of the channels were simulated using a MCFA-1 strategy, with the other six simulated using MCFA-2. MCFA-2 was chosen because it has

the poorest frequency resolution of the MCFA-N algorithms where $N > 1$, which implies that the spectral cues will be the least accurate for MCFA-2 and thereby provide the least aid in a situation in which two channels are simulated via MCFA-1. Thus, it was hypothesized that impairments were likely to have the greatest effect on the performance on MCFA-2, thereby giving a worst case response. The hypothesis to be considered is whether the additional presentation frequencies could be used to improve speech recognition when 25% of the channels were impaired.

3. Methods

3.1. Subjects and equipment

Sixteen subjects were recruited for this set of experiments. Four subjects participated in both experiments, while twelve subjects participated in only one of the experiments. A total of ten subjects completed each experiment. The subjects were recruited from the undergraduate student, graduate student and staff populations of Duke University. Four subjects had previous experience participating in acoustic model listening experiments. None of the subjects had been exposed to the speech processed by the acoustic models used in this study other than the standard model.

Prior to data collection, audiograms were obtained for the experimental subjects. Thresholds were measured at frequencies of 250, 500, 1000, 2000, 4000, and 8000 Hz to ensure that no significant hearing loss was present. Tones for the threshold measurements were presented using equipment manufactured by Tucker–Davis Technologies. A 2-down/1-up 2-interval forced choice adaptive threshold task was implemented using software obtained from Tucker–Davis. Both the threshold tasks and the speech recognition tasks were programmed and run using an IBM compatible computer in a soundproof booth (IAC, Bronx, NY). The interface for the speech recognition tasks was programmed using MATLAB[®]. All stimuli were presented binaurally to the subjects using Sony MDF-V600 dynamic stereo headphones at a listening level that was most comfortable for the subjects. Subjects responded to the task by clicking a button with a mouse. During testing procedures, feedback was not provided.

Informed consent was obtained from all subjects. Subjects were compensated for their participation. Approval was granted by the Duke University Institutional Review Board for the procedures involving human subjects as described in this paper.

3.2. Stimuli

A vowel recognition task and a consonant recognition task were used to assess speech recognition in both of the experiments. Vowel recognition was measured using a nine-choice (had, hawed, head, heard, heed, hid, hood,

hud, who'd) closed set medial vowel test (Tyler et al., 1986). Each token was presented five times in random order. A 14-choice (b, d, f, g, j, k, m n, p, s, sh, t v, z) closed set medial consonant test was also used, with consonants presented in an /aCa/ format. Again, each token was presented five times in random order. Vowel and consonant tokens were presented in noise in these experiments. Speech-shaped noise from the HINT test was utilized (Nils-son et al., 1994). Four signal-to-noise ratios (SNRs) were tested, -5 dB, 0 dB, 5 dB, and quiet. Testing always progressed from highest SNR to lowest SNR.

3.3. Experiment 1

In the first experiment, ten normal hearing subjects were tested with five different acoustic models at the four SNRs listed above. The order in which models were tested was randomized across subjects. The models tested in this first experiment were the 8-channel standard model, the modified FAME model which incorporated 2 ms analysis windows, and MCFA-2, MCFA-4, and MCFA-8. At the outset of the study, subjects were trained in quiet using the model with which they would be tested first. Both vowel and consonant tasks were used in training, and feedback was provided during the training tasks. Training was discontinued when subjects reached a 90% correct plateau on the speech tasks for the given model in quiet conditions.

Following training, subjects completed the vowel and consonant tests on each of the five models at each of the four SNRs, from highest SNR to lowest SNR. Prior to a test session with a particular model, subjects were familiarized with the processed tokens through a session that randomly presented each token twice and provided feedback. These data were not used in the data analysis. Testing consisted of five random presentations, without feedback, of each of the vowel or consonant tokens. After a phoneme was presented, the listener chose their answer by clicking the button corresponding to the desired phoneme from a closed set list.

3.4. Experiment 2

Experiment 2 mimics Experiment 1 except that a different set of models was used. Again, ten normal-hearing subjects were tested with all models. Three SNRs were tested: quiet, 0 dB and -5 dB. Experiment 2 tested both the hypothesis that MCFA-2 provides equivalent performance to doubling the number of channels of the standard model from eight to sixteen, as well as whether performance gains can still be achieved under conditions of poor frequency discrimination. The models tested included the 8-channel standard, the 16-channel standard, the 8-channel MCFA-2, and four 'impaired' 8-channel MCFA-2 models. In the four impaired models, two adjacent channels were implemented as MCFA-1, the other six were implemented as MCFA-2. The two channels that used MCFA-1 were either 1 and 2, 3 and 4, 5 and 6, or 7 and 8.

4. Results

4.1. Spectral analysis

First, in order to assess the impact of the additional frequency information, spectrograms of the speech token ‘hood’ were calculated and are illustrated in Fig. 2. Similar to the results in Nie et al. (2005), the spectrograms demonstrate that frequency transitions are lost in an 8-channel sys-

tem without adjustable presentation frequencies. However, an algorithm such as FAME retains much of the frequency transition information. With its continuum of possible frequencies, the FAME spectrogram most closely matches the original spectrogram visually; however, spectrograms that are similar to FAME and appear to contain at least partial frequency transition information can be generated with far fewer available frequencies, as demonstrated by the spectrograms associated with MCFA-8 and MCFA-4.

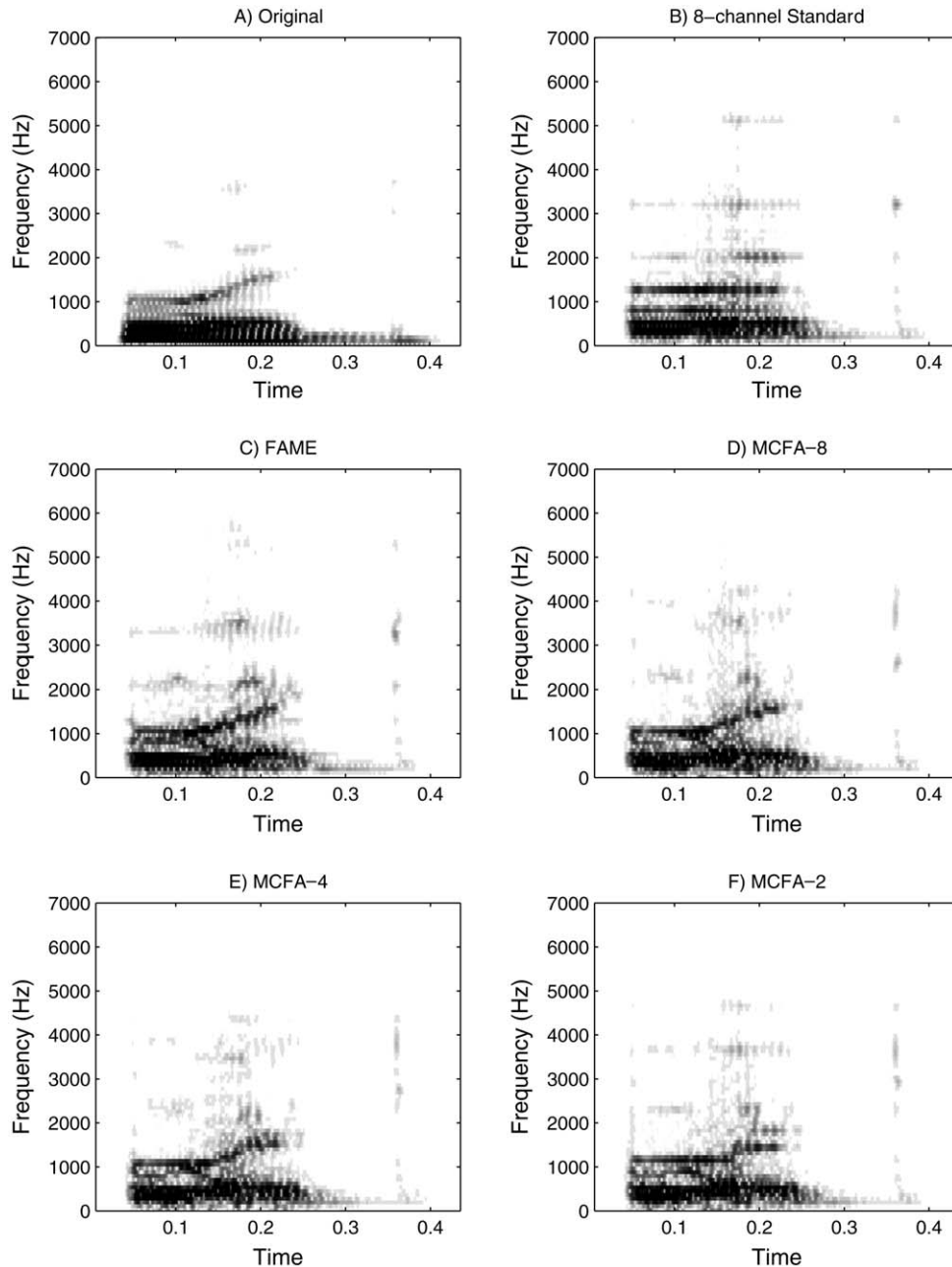


Fig. 2. The spectrograms of the speech token ‘hood’ under different conditions. (A) The original acoustic token; (B) the acoustic output of the 8-channel standard model; (C) the acoustic output of the implementation of the FAME algorithm used in this study; (D)–(F) the acoustic output of MCFA-8, MCFA-4, and MCFA-2 respectively. In (A), a frequency transition occurs from approximately 150–250 ms in the 1–2 kHz frequency range. The model spectrograms demonstrate the loss of this frequency transition information for the standard 8-channel model and the retention of that information with as few as two available presentation frequencies per channel.

Although deterioration of frequency information is visible in the spectrogram for MCFA-2, it is still capable, with just two frequencies per channel, of providing the frequency transition information. Fig. 3 demonstrates that this frequency transition information can also be retained by doubling the number of channels; however, this option may be unrealistic in implant patients who have already been implanted. The spectrograms suggest that a small increase in the encoded frequency information may retain spectral cues which would otherwise be preserved only through a large number of channels.

4.2. Experiment 1

Speech recognition for vowel and consonant tokens was measured to verify the visual observations that frequency encoding preserves useful spectral cues. The percent correct scores as a function of noise level and model from Experiment 1 are shown in Fig. 4. As expected, the FAME model results in the best performance of all the models tested in Experiment 1, though its performance was not always significantly different from MCFA-8. Statistical significance was calculated using a binomial distribution assumption

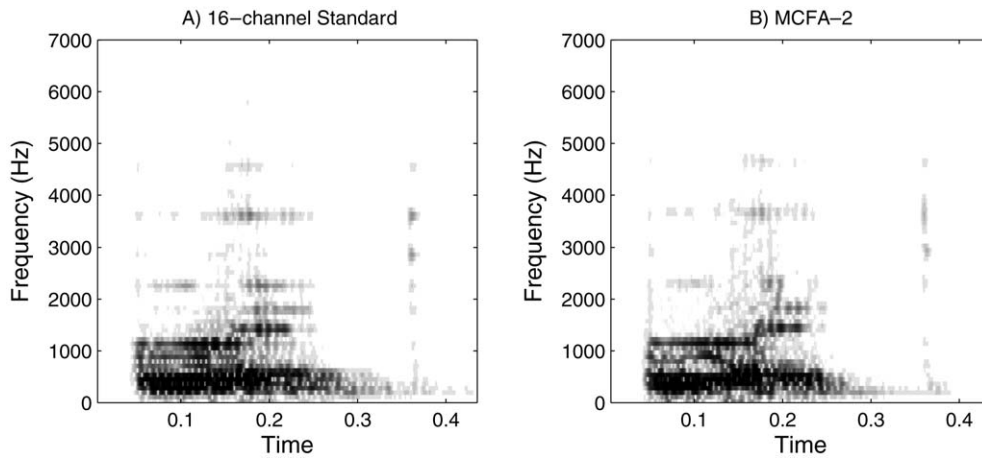


Fig. 3. The spectrograms of the speech token ‘hood’ for a 16-channel standard model (left) and MCFA-2 (right). Both spectrograms retain the frequency transition information seen in the original acoustic token spectrogram (Fig. 2(A)).

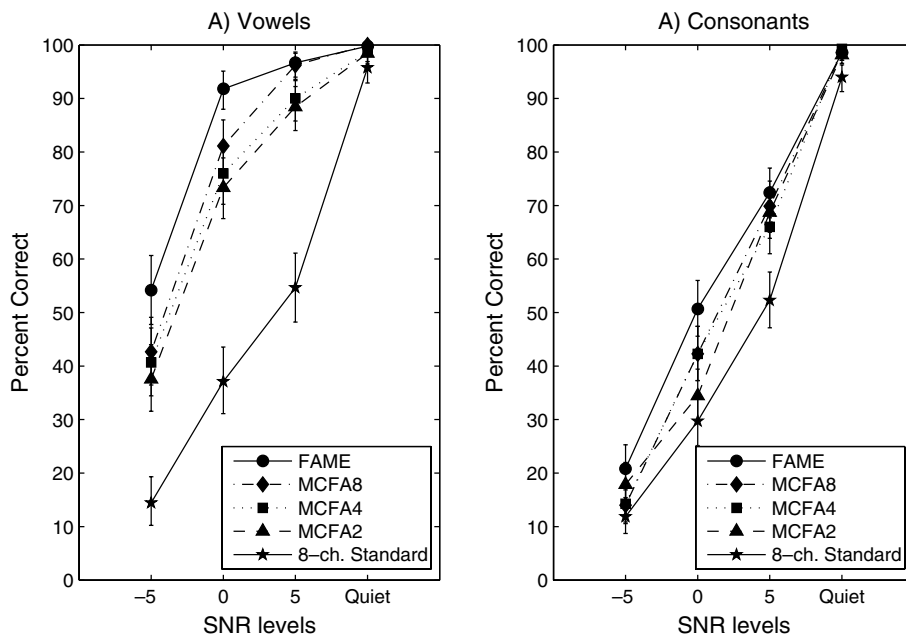


Fig. 4. Speech recognition results measured as percent correct for FAME, MCFA-8, -4, -2, and the 8-channel standard models plotted versus noise level. The left panel shows results for vowel recognition, and the right panel shows results for consonant recognition. The error bars represent the 95% confidence interval. Statistical significance ($p < 0.05$) occurs for conditions whose speech recognition scores do not have overlapping error bars. FAME provides the best performance; however, for vowel recognition, as few as two frequencies per channel provides substantial improvement over the 8-channel standard model.

(Thornton and Raffin, 1978) to account for the upper limit in speech scores.

The FAME model implemented in this study has performance levels similar to those shown by Nie et al. (2005), though the performance measured by Nie et al. at -5 dB SNR was greater by approximately 30 percentage points for both vowels and consonants. This difference may be due to implementation differences such as holding the frequency constant over each 2 ms analysis window or procedural/training differences. Subjects in the current study were trained in quiet until performance reached a plateau, and feedback was not provided during testing. At the other SNRs, the performance levels measured for FAME match those of Nie et al. (2005) within 10 percentage points or fewer.

For vowel recognition and low SNRs, FAME achieved significantly better performance than the MCFA models ($p < 0.05$); however, for consonant recognition this was not the case. Thus, under some circumstances, a small increase in the available presentation frequencies may provide a similar level of benefit as a continuum of presentation frequencies. The differences between the scores of the MCFA models were not statistically significant; however, for vowels in noise, the MCFA models did provide substantial, significant improvement ($p < 0.05$) over the standard model. This includes MCFA-2 with just an increase of one additional frequency available per channel. For consonants, significant improvement over the standard model ($p < 0.05$) required at least four frequencies per channel, and the improvements in recognition were much lower

than for vowel recognition. The results of Experiment 1 are promising for implementation in implant subjects since they suggest that even small increases in the amount of encoded frequency information have the potential for significant improvement over the performance of an equivalent algorithm that does not encode frequency. However, acoustic models only provide a trend in performance, not the magnitude of a possible performance gain, thus the degree to which performance could be improved in implant subjects remains to be seen.

4.3. Experiment 2

The relative importance of encoding frequency is demonstrated by comparing the speech recognition performance of MCFA-2 to that of a 16-channel version of the standard model. The 16-channel standard model presents twice as many carrier signals as the 8-channel MCFA-2 model; however, the MCFA-2 model has a choice of eight out of sixteen available frequencies. As shown in Fig. 5, the percent correct results as a function of SNR and model demonstrate that the differences in performance between MCFA-2 and the 16-channel standard model were not significantly different. This agrees with the similarity observed between the spectrograms of the two models as discussed previously. Thus, providing the addition of a small level of encoded frequency information has the potential to attain speech recognition performance equivalent to doubling the number of channels, assuming that this information can be conveyed to implant subjects adequately.

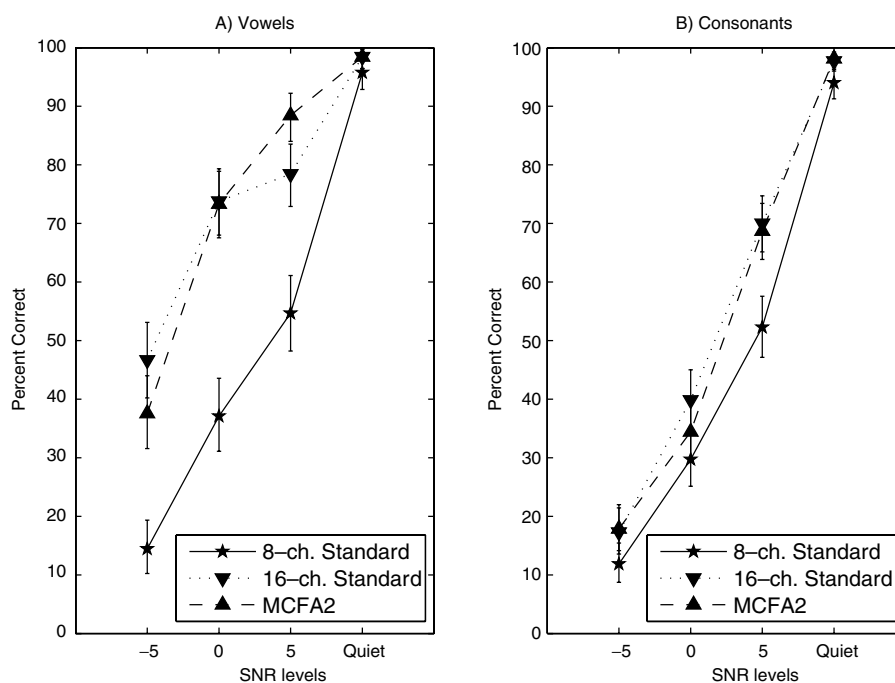


Fig. 5. Comparison of speech recognition scores measured as percent correct for the MCFA-2 and 16-channel standard models plotted versus noise level. The left panel shows results for vowel recognition, and the right panel shows results for consonant recognition. The error bars represent the 95% confidence interval. Statistical significance ($p < 0.05$) occurs for conditions whose speech recognition scores do not have overlapping error bars. Performance differences between the two models are not statistically significant.

Although the performance results from the MCFA models are promising, the MCFA algorithm (and the FAME algorithm) inherently assumes that all presentation frequencies are perfectly discriminable. This may not be the case for implant patients each of whom may have unique limitations. Experiment 2 considers the effects of poor frequency discrimination on the ability of MCFA-2 to provide improvements in speech recognition (see Fig. 6). For vowel recognition, poor frequency discrimination on the central channels (3 and 4, or 5 and 6) resulted in performance that did not differ significantly from the performance of the 8-channel standard model. However, poor frequency discrimination on the outer channels still resulted in significant performance improvements ($p < 0.05$). Thus, poor frequency discrimination, localized to the extremes of the set of channels do not appear to degrade the performance improvements provided by MCFA for vowel recognition. Little can be determined from the consonant recognition results since MCFA-2 does not provide statistically significant performance improvements even without impairment.

The patterns in the vowel recognition results were further investigated using formant information transmission analysis (Miller and Nicely, 1955). Miller and Nicely (1955) proposed grouping phonemes by the features that they share such that confusions between phonemes within

a group would not be considered incorrect responses. Once the phonemes were grouped by feature, the entropy of the stimuli and the corresponding responses is used to calculate a transmission rate. Fig. 7 shows the results for the first and second formants (F1 and F2, respectively) for the four impaired models. For transmission of both F1 and F2, the models that impair the outer channels have the highest performance due to the low amount of formant information contained in these channels. However, the transmission of F1, as seen in the left panel, is lowest for impairment of channels 3 and 4 where the most first formant information is located. Similarly, in the right panel, transmission of F2 is lowest for channels 5 and 6. Thus, the impairment on the central channels leads to reduced formant transmission which likely then results in the lowered vowel recognition scores.

5. Discussion

As demonstrated in the spectrograms shown in Figs. 2 and 3, degradation of spectral cues can occur for low levels of frequency information as is the case with processors/systems with few channels and no frequency encoding. This in turn may result in poorer speech recognition. The spectrograms suggest that these cues may be retained by increased frequency resolution, either through an increase in the

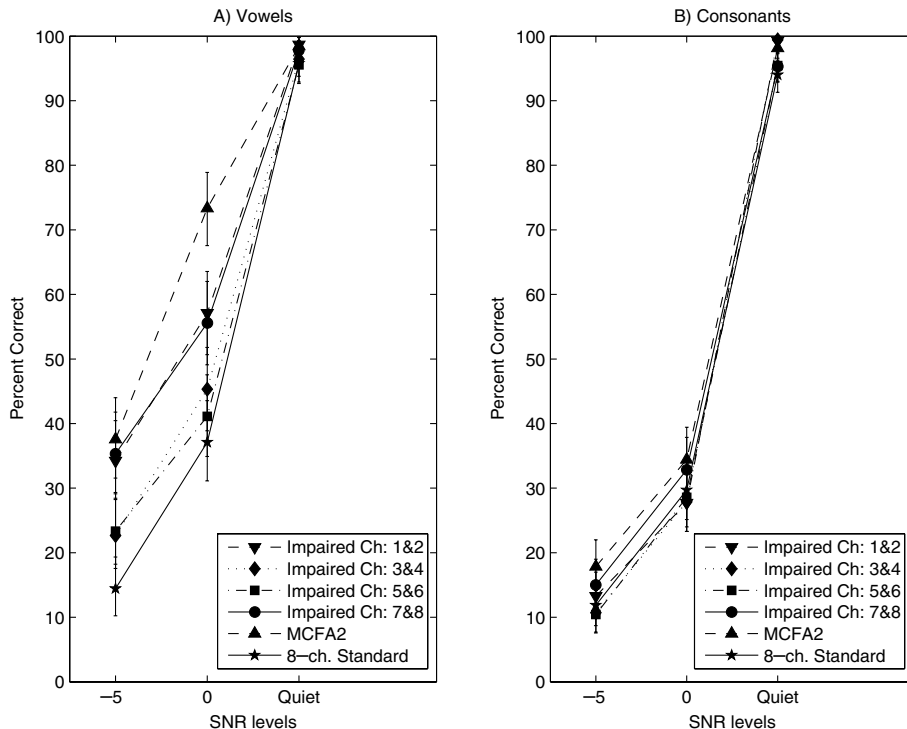


Fig. 6. Speech recognition results measured as percent correct for MCFA-2, the 8-channel standard, and the four impaired MCFA-2 models plotted versus noise level. The left panel shows results for vowel recognition, and the right panel shows results for consonant recognition. The error bars represent the 95% confidence interval. Statistical significance ($p < 0.05$) occurs for conditions whose speech recognition scores do not have overlapping error bars. For vowel recognition, the differences in performance between the 8-channel standard model and the models that impaired the central channels (3 and 4, and 5 and 6) were not statistically significant. However, the other two models, despite the impairments on the outer channels, did have statistically significant performance improvements over the 8-channel standard model. For consonant recognition, none of the MCFA models had speech recognition scores significantly different than the 8-channel standard model.

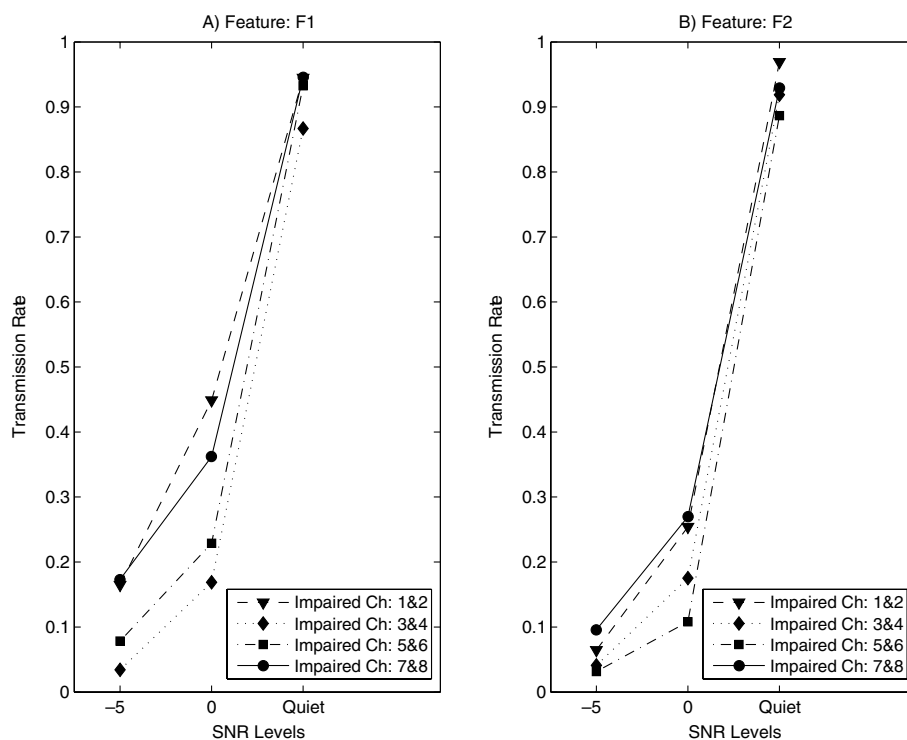


Fig. 7. Information transmission rates of the first (left panel) and second (right panel) formants for the four impaired MCFA models as a function of SNR. Models that impair the outer channels consistently transmit more formant information than the models that impair central channels. For the first formant, impairing the lower-frequency central channels (3 and 4) results in the lowest transmission. Similarly, for the second formant, impairing the higher-frequency central channels (5 and 6) results in the lowest transmission.

number of channels or through frequency modulation within channel. Speech recognition scores verified that increasing frequency resolution provides significant performance improvement. Further, increasing the number of available presentation frequencies by a single frequency per channel was equivalent to doubling the number of channels. Increasing the number of channels may be an unrealistic approach to improving speech recognition since cochlear implant subjects may already be limited by the number of channels available through their current implanted electrode array. Further, increasing the number of channels might lead to increased channel interactions which have been shown in some cases to be correlated with decreased speech recognition (e.g. Throckmorton and Collins, 1999; Henry et al., 2000). Thus, a method that can achieve the performance improvements characteristic of increasing the number of channels without increasing the number of channel interactions may be preferable.

Some results by Zeng et al. (2005) that compare models that increase frequency resolution through frequency modulation within channel rather than by increasing the number of channels also suggest that increasing channels is not necessarily the optimal method of providing more frequency resolution. While the performance for a 16-channel amplitude modulation model matched the performance of an 8-channel model with frequency modulation within channel for speech recognition in noise, Zeng et al. (2005) demonstrated that this may not be the case for other types

of tasks such as speaker recognition and Mandarin tone recognition.

As mentioned in the Introduction, acoustic models are only intended to demonstrate trends in performance rather than make exact predictions of performance. This is further influenced in this study by the difference between the implementation in acoustic models and the intended implementation in implant subjects. For acoustic models, the pitch percept is tied to frequency; however, in implant subjects, the ability to generate a pitch percept equivalent to a given frequency is not possible due to an inability to compare electric and acoustic stimulation. Pitch percepts for implant subjects will be chosen based on which rates produce detectable pitch changes. Thus, it is possible that acoustic model results are inflated due to an exact match between pitch and frequency which will most likely not be possible in the implant implementation.

However, Fearn (2001) did demonstrate an increase in music appreciation in implant subjects when pulse rates were used to increase frequency resolution within channel. These results would suggest that frequency modulation within channel may have benefits not produced by merely increasing the number of channels, and the results of Fearn (2001) suggest that this frequency modulation may be possible through variable pulse rate stimulation.

However, encoding frequency through stimulation rate may be more difficult to implement than increasing the number of channels, when increasing channels is possible.

Although the STFT provides a computationally fast and simple method of estimating frequency, its resolution is limited by the sample duration, which is for this acoustic model, and presumably will be for implants, quite short – on the order of milliseconds. The short sample duration limits the resolution of individual peaks in the frequency domain; however, it may not necessarily invalidate this approach since the goal is to focus the frequency estimation to an appropriate area of the domain rather than to calculate an accurate representation of the spectrum. Especially in the case of MCFA-2, the interest is whether most of the energy is in the lower or higher section of the passband. To effectively utilize the greater number of frequencies provided by MCFA-4 and -8, a greater frequency resolution is probably necessary, and this may explain in part why these two models did not provide significant improvement over MCFA-2. The poor frequency resolution might also explain why impairing the lowest channels (where the small passbands cause the poor resolution to have the greatest effect) did not significantly lower performance. Possibly those channels were already impaired by the poor resolution.

For future implementation, consideration must be given as to whether the computational simplicity of the STFT might be better sacrificed for a more accurate measure of frequency. Alternatively, the STFT could be updated less frequently than the amplitude modulation, thereby increasing the sample duration. Either method may provide an increase in the accuracy of the frequency mapping. However, initial results with an acoustic model (Kucukoglu and Collins, 2006) suggest that at least with normal-hearing subjects, frequency extraction accuracy does not have a significant effect unless the noise level is quite high (SNR = -5 dB). With implementation in implant subjects, there may be a threshold of accuracy beyond which little to no increase in performance is gained, i.e. the limitation in frequency perception comes not from the frequency estimates calculated by the algorithm but from the limitation of pulse rates to evoke that perception. While general trends in frequency may be mapped to pulse rate, exact representations of frequency are unlikely to be attainable through this algorithm. So, increasing the computational complexity in order to increase the accuracy of the frequency estimation may eventually cease to provide benefits.

Aside from algorithm design, other issues may complicate implementation of a multiple-rate signal processing strategy. The range of pulse rates that might be effective is generally different than the range of pulse rates that are actually available from the device. Thus, psychophysical investigation to determine the effective set of pulse rates will be necessary to optimize the performance gain. Further, the range of effective pulse rates is likely to be subject dependent. Cochlear implant subjects, due to unknown physiology, may have poor pulse rate resolution, and if pulse rate is used to generate different frequency percepts, then poor resolution in rate will lead to poor resolution in frequency. Experiment 2 demonstrated that if poor fre-

quency resolution occurs in channels that correspond to important cues, then the benefit gained by modulating frequency may be negated. For these two reasons, implementation will likely require tuning for individual subjects.

Although Experiment 2 does suggest that in some cases poor frequency discrimination entirely negates the benefit of frequency modulation within channel, Experiment 2 presents a worst case in which no amount of frequency variation can create separate pitch percepts. Since frequent investigation into the effect of pulse rate on pitch has failed to discover a case in which pulse rate had no effect whatsoever (e.g. Shannon, 1983; Tong et al., 1983; Townshend et al., 1987; Zeng, 2002; Landsberger and McKay, 2005), the ability to generate variable pitch percepts through variable pulse rates does not seem in doubt; however, whether implementation can be carried out such that implant subjects can use this information in a meaningful manner remains to be determined. Experiment 2 could also be considered a representation of the effects of tuning the modulation frequencies poorly, such that two different frequencies result in the same pitch percept. Such a possibility again emphasizes that one important aspect of implementation will be tuning the multirate algorithm for individual subjects.

Another issue that will require further investigation is the perceptual overlap across channels that may occur with a multirate algorithm. At best, perceptual overlap might negate the possible benefits of a multirate system, but at worst, perceptual overlap might create psychophysical anomalies, such as pitch reversals, that have been suggested as factors capable of significantly degrading speech recognition (Throckmorton and Collins, 2002). Recent studies of the effect on pitch of varying pulse rates, compared across electrodes, suggest that a significant amount of pitch overlap may occur (Fearn and Wolfe, 2000; Zeng, 2002). However, these studies used pitch scaling tasks to determine pitch overlap, which may not be an adequate measure. In testing pitch scaling across electrodes stimulated under identical conditions, Collins et al. (1997) observed a standard deviation of 10% of the scale, suggesting an inherent difficulty in accurate scaling of pitch. Further, for Zeng (2002), three out of four subjects estimated the pitches of the lowest pulse rate presented on an apical and a basal electrode to be equal, thereby using almost the entire scale (0–100) for both electrodes. These results seem to suggest that, at least in part, subjects may have split the pitch scaling task into two tasks, estimating pitch for each electrode separately.

Pitch ranking, although not providing scale, may be a preferable task for determining pitch overlap since subjects are not required to remember the entire set of stimuli in order to make a judgment. Pitch ranking does assume a one-dimensional pitch percept despite a combination of both stimulation rate and electrode place, and this may not necessarily be the case. Several studies have indicated that place and rate pitch are separate perceptual dimensions for implant subjects (McKay et al., 2000; Tong

et al., 1983). Research will be needed to determine the extent that perceptual dimensions affects perceptual overlap. Overlap might potentially be eliminated if the distinction between combinations of rate and place is large. On the other hand, this separation is purely a construct of electrical stimulation, and whether cochlear implant subjects can utilize these separate forms of information to experience a greater number of pitch percepts is as yet unknown.

In addition to the interaction between rate and place of stimulation, it may be necessary to determine whether changing pulse rate changes perception of amplitude modulation. McKay and McDermott (1996) determined that implant subjects perceive an aggregate temporal pattern for electrodes that are closely spaced, but not for those with greater separation, when pulse rate is kept constant. It is possible that this perceptual characteristic might change under variable pulse rates; the effects of which on speech recognition would need to be considered.

Thus, although this modeling study suggests that providing additional frequency resolution through frequency modulation within channel may provide a significant increase in speech recognition, even for small increases in spectral information such as in the case of MCFA-2, several issues will need to be investigated in order to effectively implement the algorithm in implant subjects. Other spectral estimation techniques need to be investigated to determine whether increased frequency resolution further improves the performance of the MCFA algorithms. Beyond the basic algorithm design, psychophysical effects will need to be considered, both in terms of effective pitch rates (perceivable and non-confounding) and individual characteristics of subjects. However, despite these caveats, the results of this study and others suggest that frequency modulation within channel may be able to provide some improvement in speech recognition performance. Further, in this study, the performance of MCFA-2 suggests that the amount of information that may need to be encoded for substantive performance improvements might be modest – an increase from just one modulation frequency per channel (MCFA-1) to two (MCFA-2) resulted in a larger increase in performance than was obtained for any further additions to the number of modulation frequencies per channel. If an effective method can be determined for increasing frequency resolution within channel in implants, these results suggest an improvement in speech recognition may be possible.

Acknowledgements

We express our thanks and appreciation to the subjects who gave us their time and effort. We also thank Ms. Danielle Davidian for her aid in collecting pilot data for this study, Dr. Fan-Gang Zeng for useful suggestions regarding this work, and two anonymous reviewers for their aid with revisions. This research was supported in part by NSF under Grant No. NSF-BES-00-85370 and NIH Grant No. R01 DC-007994-01.

References

- Baskent, D., Shannon, R.V., 2003. Speech recognition under conditions of frequency-place compression and expansion. *Journal of the Acoustical Society of America* 113 (4), 2064–2076.
- Blamey, P.J., Dowell, R.C., Clark, G.M., 1987. Acoustic parameters measured by a formant-estimating speech processor for a multiple-channel cochlear implant. *Journal of the Acoustical Society of America* 82 (1), 38–47.
- Clark, G.M., 1987. The University of Melbourne-Nucleus multi-electrode cochlear implant. *Advances in Oto-Rhino-Laryngology* 38, 1–189.
- Collins, L.M., Zwolan, T.A., Wakefield, G.H., 1997. Comparison of electrode discriminability, pitch ranking, and pitch scaling data in postlingually deafened adult cochlear implant subjects. *Journal of the Acoustical Society of America* 101 (1), 440–455.
- Dorman, M.F., Loizou, P.C., Fitzke, J., Tu, Z., 1998. The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels. *Journal of the Acoustical Society of America* 104, 3583–3585.
- Dorman, M.F., Loizou, P.C., Kemp, L.L., Kirk, K.I., 2000. Word recognition by children listening to speech processed into a small number of channels: Data from normal-hearing children and children with cochlear implants. *Ear & Hearing* 21 (6), 590–596.
- 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.
- Dorman, M.F., Loizou, P.C., Spahr, A.J., Maloff, E., 2002a. A comparison of the speech understanding provided by acoustic models of fixed-channel and channel-picking signal processors for cochlear implants. *Journal of Speech, Language, and Hearing Research* 45, 783–788.
- Dorman, M.F., Loizou, P.C., Spahr, A.J., Maloff, E., 2002b. Factors that allow a high level of speech understanding by patients fit with cochlear implants. *American Journal of Audiology* 11, 119–123.
- Fearn, R., 2001. Music and pitch perception of cochlear implant recipients. Dissertation. University of New South Wales, Sydney, Australia.
- Fearn, R., Wolfe, J., 2000. Relative importance of rate and place: Experiments using pitch scaling techniques with cochlear implant recipients. *Annals of Otology, Rhinology, and Laryngology – Supplement* 109 (12), 51–53.
- Flanagan, J.L., 1980. Parametric coding of speech spectra. *Journal of the Acoustical Society of America* 68, 412–419.
- Flanagan, J.L., Golden, R.M., 1966. Phase vocoder. *Bell System Technical Journal* 45, 1493–1509.
- Friesen, L.M., Shannon, R.V., Baskent, D., Wang, X., 2001. Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *Journal of the Acoustical Society of America* 110 (2), 1150–1163.
- Fu, Q.-J., Shannon, R.V., 1999. Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing. *Journal of the Acoustical Society of America* 105, 1889–1900.
- Fu, Q.-J., Shannon, R.V., Wang, X., 1998. Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing. *Journal of the Acoustical Society of America* 104, 3586–3596.
- Geier, L., Fisher, L., Barker, M., Opie, J., 1999. The effect of long-term deafness on speech recognition in postlingually deafened adult Clarion cochlear implant users. *Annals of Otology, Rhinology, and Laryngology – Supplement* 177, 80–83.
- Hartmann, R., Topp, G., Klinke, R., 1984. Discharge patterns of cat primary auditory fibers with electrical stimulation of the cochlea. *Hearing Research* 13, 47–62.
- Henry, B.A., McKay, C.M., McDermott, H.J., Clark, G.M., 2000. The relationship between speech perception and electrode discrimination in cochlear implantees. *Journal of the Acoustical Society of America* 108, 1269–1280.

- Kessler, D.K., 1999. The Clarion multi-strategy cochlear implant. *Annals of Otolaryngology, Rhinology, and Laryngology – Supplement* 177, 8–16.
- Kiang, N.Y.-S., Moxon, E.C., 1972. Physiological considerations in artificial stimulation of the inner ear. *Annals of Otolaryngology, Rhinology, and Laryngology* 81, 714–730.
- Kucukoglu, M.S., Collins, L.M., 2006. Alternative methods for frequency selection in a frequency encoding algorithm for cochlear implant speech processors. 29th ARO Midwinter Meeting, Baltimore, MD.
- Landsberger, D.M., McKay, C.M., 2005. Perceptual differences between low and high rates of stimulation on single electrodes for cochlear implantees. *Journal of the Acoustical Society of America* 117 (1), 319–327.
- Loizou, P.C., Dorman, M.F., Poroy, O., Spahr, A.J., 2000. Speech recognition by normal-hearing and cochlear implant listeners as a function of intensity resolution. *Journal of the Acoustical Society of America* 108 (5), 2377–2387.
- McKay, C.M., McDermott, H.J., 1996. The perception of temporal patterns for electrical stimulation presented at one or two intracochlear sites. *Journal of the Acoustical Society of America* 100 (2), 1081–1092.
- McKay, C.M., McDermott, H.J., Carlyon, R.P., 2000. Place and temporal cues in pitch perception: Are they truly independent? *Acoustics Research Letters Online* 1 (1), 25–30.
- McKay, C.M., McDermott, H.J., Vandali, A.E., Clark, G.M., 1992. A comparison of speech perception of cochlear implantees using the Spectral Maxima Sound Processor (SMSP) and the MSP (MULTI-PEAK) processor. *Acta Otolaryngologica* 112, 752–761.
- Miller, G.A., Nicely, P.E., 1955. An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America* 27, 338–352.
- Nie, K., Stickney, G., Zeng, F.-G., 2005. Encoding frequency modulation to improve cochlear implant performance in noise. *IEEE Transactions on Biomedical Engineering* 52 (1), 64–73.
- Nilsson, M., Soli, S.D., Sullivan, J.A., 1994. Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America* 95 (2), 1085–1099.
- Patrick, J.F., Clark, G.M., 1991. The Nucleus 22-channel cochlear implant system. *Ear and Hearing – Supplement* 12 (4), 3S–9S.
- Rubinstein, J.T., Parkinson, W.S., Tyler, R.S., Gantz, B.J., 1999. Residual speech recognition and cochlear implant performance effects of implantation criteria. *American Journal of Otolaryngology* 20, 445–452.
- Shannon, R.V., 1983. Multichannel electrical stimulation of the auditory nerve in man. I. Basic psychophysics. *Hearing Research* 11, 157–189.
- Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. *Science* 270, 303–304.
- Shipp, D.B., Nedzelski, J.M., 1995. Prognostic indicators of speech recognition performance in adult cochlear implant users: a prospective analysis. *Annals of Otolaryngology, Rhinology, and Laryngology – Supplement* 166, 194–196.
- Thornton, A.R., Raffin, M.J.M., 1978. Speech discrimination scores modeled as a binomial variable. *Journal of Speech and Hearing Research* 21, 507–518.
- Throckmorton, C.S., Collins, L.M., 1999. Investigation of the effects of temporal and spatial interactions on speech-recognition skills in cochlear-implant subjects. *Journal of the Acoustical Society of America* 105, 861–873.
- 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 (1), 285–296.
- Tong, Y.C., Blamey, P.J., Dowell, R.C., Clark, G.M., 1983. Psychophysical studies evaluating the feasibility of a speech processing strategy for a multiple-channel cochlear implant. *Journal of the Acoustical Society of America* 74 (1), 73–80.
- Townshend, B., Cotter, N., Van Compernelle, D., White, R.L., 1987. Pitch perception by cochlear implant subjects. *Journal of the Acoustical Society of America* 82 (1), 106–115.
- Tyler, R.S., Preece, J.P., Tye-Murray, N., 1986. The Iowa Phoneme and Sentence Tests. Iowa City, The University of Iowa, Department of Otolaryngology.
- van Dijk, J.E., van Olphen, A.F., Langereis, M.C., Mens, L.H.M., Brokx, J.P.L., Smoorenburg, G.F., 1999. Predictors of cochlear implant performance. *Audiology* 38, 109–116.
- Wilson, B.S., Lawson, D.T., Finley, C.C., Wolford, R.D., 1991. Coding strategies for multichannel cochlear prostheses. *The American Journal of Otolaryngology – Supplement* 12, 56–61.
- Zeng, F.-G., 2002. Temporal pitch in electric hearing. *Hearing Research* 174, 101–106.
- Zeng, F.-G., Nie, K., Stickney, G., Kong, Y.Y., Vongphoe, M.A.B., Weit, C.G., Cao, K., 2005. Speech recognition with amplitude and frequency modulations. *Proceedings of the National Academy of Sciences* 102 (7), 2293–2298.
- Zwolan, T.A., Collins, L.M., Wakefield, G.H., 1997. Electrode discrimination and speech recognition in postlingually deafened adult cochlear implant subjects. *Journal of the Acoustical Society of America* 102 (6), 3673–3685.