IMPORTANCE OF PITCH AND PERIODICITY TO CHINESE-SPEAKING COCHLEAR IMPLANT PATIENTS

Xin Luo and Qian-Jie Fu

Department of Auditory Implants and Perception
House Ear Institute, 2100 West Third Street, Los Angeles, CA 90057, USA
Email: xluo@hei.org

ABSTRACT

This study examines the relative importance of pitch and periodicity to Chinese speech recognition by normal-hearing subjects listening to a cochlear implant simulation. Three carrier band conditions were tested, each of which provided different amounts of pitch and periodicity information: 1) noise-band carriers for all speech segments, 2) pulse train carriers for voiced speech segments, in which the rate followed the fundamental frequency (*F*0) of the speech signal, and 3) fixed-rate (150 Hz) pulse train carriers for voiced speech segments. The noise-band carriers preserved little pitch and periodicity information, as all temporal cues were limited by the 50-Hz temporal envelope extracted from each frequency band. The *F*0-controlled pulse train carriers preserved all pitch and periodicity information. The fixed-rate pulse train carriers preserved periodicity information, but no pitch information. Results showed that different carriers produced significantly different amounts of Chinese speech recognition, with the *F*0-controlled pulse train carriers producing the best performance. These results reflect the need to deliver adequate amounts of both pitch and periodicity information to Chinese-speaking cochlear implant patients.

1. INTRODUCTION

The pitch and periodicity information of the glottal wave during speech production is important for speech understanding, especially for tonal languages such as Chinese [1]. For example, the pitch contour allows for identification of speakers, as well as speakers' intonation, while periodicity information is essential for distinguishing voiced and unvoiced speech segments.

Cochlear implants (CIs) represent speech signals by using the temporal envelope extracted from frequency analysis bands to modulate pulse trains delivered to appropriate implanted electrodes. Typically, there are 8 to 24 implanted electrodes, which severely limits the spectral resolution available to CI patients. Many researchers have studied the perception of pitch and periodicity information using CI simulations with normal-hearing (NH) listeners, thereby measuring the importance of pitch and periodicity cues to speech recognition under conditions of reduced spectral resolution. For English speech perception, carriers with different amounts of pitch and periodicity information (noise-band, fixed-rate pulse train, and *F*0-controlled pulse train) all produced similar levels of recognition performance, indicating that, for English, the contribution of pitch and periodicity cues was relatively weak [2].

In contrast to English, Chinese is a tonal language in which pitch patterns of vowels are lexically important. Studies have demonstrated the importance of tonal envelope cues in Chinese speech recognition by CI users [3]. Different amounts of pitch and periodicity cues may produce different levels of Chinese speech recognition in CI.

In the present study, the effects of pitch and periodicity information on Chinese speech recognition were measured for normal-hearing native Chinese subjects listening to a cochlear implant simulation. Recognition of Chinese vowels, consonants, tones, and sentences was compared to English speech recognition under similar speech processing conditions. The importance of pitch and periodicity information to Chinese speech recognition by CI users was analyzed using Boothroyd's model [4].

2. EXPERIMENT DESIGN

2.1. Subjects

Six young adult native Chinese-speaking listeners (3 males, 3 females) participated in the study. All subjects were normal hearing and had pure-tone thresholds better than 20 dB HL at octave frequencies from 125 Hz to 8000 Hz in both ears.

2.2. Stimuli and Speech Processing

Speech test materials in the present study included Chinese vowel, consonant, and sentence stimuli. Chinese vowel and consonant stimuli were derived from the 'Chinese Standard Database' [5]. Five male and five female speakers each produced 4 tones for 6 Mandarin Chinese single-vowel syllables (/a/, /o/, /e/, /i/, /u/, /ü/), resulting in a

total of 240 vowel tokens. These vowel stimuli were used for measuring both Chinese vowel and tone recognition. Using the 21 Chinese initial consonants used in the consonant recognition tests, one male and one female speakers each produced 4 tones for /u/ in a consonant-/u/ context, thereby creating a set of 152 lexically meaningful combinations. Chinese sentence stimuli were derived from the Mandarin Hearing in Noise Test (HINT) sentences [6]. One male speaker produced 240 Chinese sentences of easy to moderate difficulty; the length of the sentences was fixed to be 10 words. The Chinese vowel and consonant stimuli were digitized using a 16-bit A/D converter at a 16-kHz sampling rate, while the Chinese sentence stimuli were sampled at a 24-kHz sampling rate.

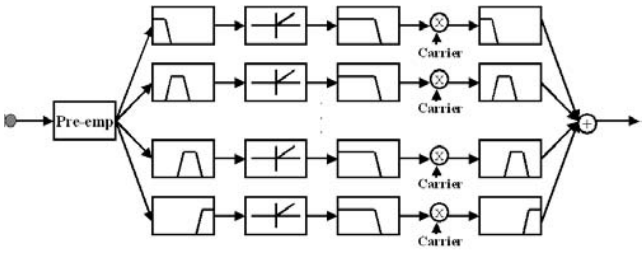

Figure 1: Block diagram of the CIS simulation.

Figure 1 shows a block diagram of the speech processor designed to simulate a cochlear implant fitted with the continuous-interleaved-sampling (CIS) strategy [7]. After pre-emphasis (1^{st}-order Butterworth high-pass filter at 1200 Hz), the input speech signal was divided into either 2 or 4 frequency bands (depending on the test condition); the overall input frequency range was 100 - 6000 Hz. The corner frequencies of the analysis bands were determined according to Greenwood's formula [8]; all analysis filters were 4^{th}-order Butterworth band-pass filters. The temporal envelope from each analysis band was extracted by half-wave rectification and low-pass filtering (4^{th}-order Butterworth low-pass filter at 50 Hz), and was used to modulate one of the three experimental carriers. The modulated carriers were band-pass filtered by filters with the same pass-bands as the analysis filters. The output speech was the sum of these band-limited, modulated carriers.

Three experimental carriers were used in the simulation, each of which provided different amounts of pitch salience and periodicity information. For the noise-band (NB) carrier, pitch and periodicity information was only available from the 50-Hz temporal envelope of each frequency band. For the *F*0-controlled pulse train (*F*0c) carrier, a pulse train in which the rate followed the fundamental frequency (*F*0) was used for voiced speech segments, and random noise was used for unvoiced speech segments. The *F*0c carrier preserved all pitch and periodicity information. For the fixed-rate pulse train (FR) carrier, a 150 Hz pulse train was used for voiced speech segments, and random noise was used for unvoiced speech segments. In the FR carrier, only periodicity information was preserved; pitch information was not preserved. Both the *F*0c and the FR carriers may be considered as "binary" carriers, in which voiced speech segments were represented by periodic carriers and unvoiced segments by random noise. Both binary carriers used mono-phasic pulses (pulse width = 1/sampling rate of stimulus); the pulse amplitudes were equal to the root mean square (RMS) levels of the random noise carriers.

The pitch and periodicity information used to control the carriers was extracted from original speech signal using an auto-correlation method on a frame-by-frame basis. The *F*0 extraction method was similar to an algorithm proposed by Markel [9], with some simplifications. To remove the influence of formant frequencies on *F*0 extraction, 12^{th}-order linear predictive (LP) analysis was performed for each frame using the Levinson-Durbin algorithm. The speech signal was inverse-filtered to obtain the prediction residual, a signal with an approximately flat spectrum. The auto-correlation of the residual signal was calculated and the location of the auto-correlation peak within an appropriate range (2 - 20 ms) was chosen as the pitch period of the frame. The Voiced/Unvoiced (VUV) quality of each frame was considered to be voiced if the normalized level of the auto-correlation peak was beyond an empirical threshold (0.2 in this study); otherwise, it was considered to be unvoiced. The *F*0 analysis frame size was adaptive; if the normal 30 ms frame size was less than 3 times of the immediately previous pitch period estimate, the analysis frame size was increased to be 4 times of the immediately previous pitch period estimate. The analysis frame shift was fixed to be 10 ms.

Several post-processing methods derived from the physiological constraints on *F*0 variations were employed to fix errors such as VUV confusion, pitch doubling, and pitch halving. For example, if a voiced frame was found to be between two unvoiced frames (or vice-versa), the VUV decision of the intermediate frame was reversed. Similarly, when pitch doubling or pitch halving occurred for two successive voiced frames, pitch values were corrected to be the one with higher voicing degree. Finally, a 5-point median filter was used to smooth the extracted *F*0 contour.

2.3. Procedure

Closed-set identification tasks were used to measure Chinese tone (4-choices), vowel (6-choices), and consonant (21-choices) recognition. Chinese sentence recognition was measured using an open-set paradigm. For each recognition task, the test order of speech processing conditions was randomized and counterbalanced across subjects. No feedback was provided for all tests. Subjects were seated in a double-walled sound-treated booth and listened to the stimuli presented in free field over a single loudspeaker (Tannoy Reveal) at 65 dBA.

3. RESULTS

Figure 2 shows Chinese speech recognition scores obtained with the three experimental carriers, as a function of the number of frequency bands (hereafter, the number of channels). Panels A, B, C, and D correspond to consonant, vowel, tone, and sentence recognition, respectively.

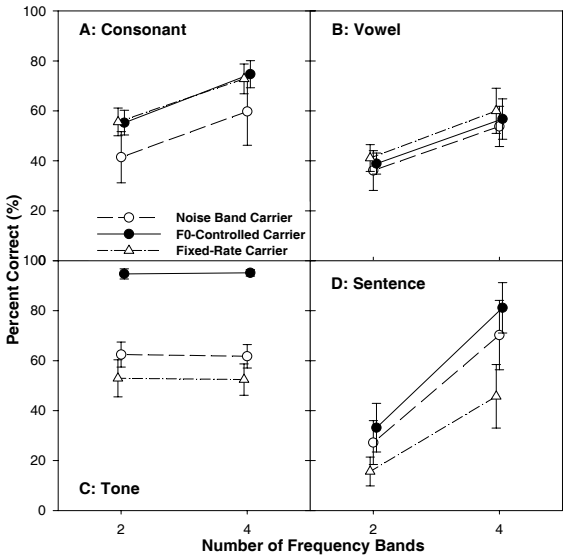

Figure 2: Chinese speech recognition with experimental carriers, as a function of the number of frequency bands.

Figure 2.A shows Chinese consonant recognition scores obtained with different carriers, as a function of the number of channels. Mean consonant recognition significantly improved for all carriers as the number of channels was increased from 2 to 4. Mean consonant recognition also significantly improved when the binary carriers (*F*0c and FR) were utilized, rather than the NB carrier; performance improved from 41 to 55 % correct with 2 channels, and from 60 to 73 % correct with 4 channels. A two-way analysis of variance (ANOVA) revealed that both the number of channels [$F(1,30)=44.24$, $p<0.001$] and the carrier type [$F(2,30)=11.49$, $p<0.001$] significantly affected Chinese consonant recognition. Significant differences in performance were found between the noise band carrier and the binary carriers, but not between the *F*0c and FR carriers [$F(1,20)=0.12$, $p=0.74$].

The amount of consonant features transmitted by each experimental processor was calculated; three production-based features of voice, manner, and place of articulation were analyzed [10]. The transmission of all features was enhanced by either increasing the number of channels, or by changing from noise-band to binary carriers. However, there was no significant difference between the amount of feature information transmitted by the *F*0c and FR carriers.

Figure 2.B shows Chinese vowel recognition scores. Vowel recognition significantly improved for all carriers as the number of channels was increased; performance improved from 38 % correct with 2 channels to 56 % correct with 4 channels. However, vowel recognition was not significantly affected by the carrier type. A two-way ANOVA showed that Chinese vowel recognition was significantly dependent on the number of channels [$F(1,30)=55.05$, $p<0.001$], but not on the carrier type [$F(2,30)=1.77$, $p=0.19$].

Figure 2.C shows Chinese tone recognition scores. In contrast to Chinese vowel recognition (Figure 2.B), mean tone recognition was not significantly affected by the number of channels; however, the carrier type greatly affected tone recognition. The FR carrier produced the lowest tone recognition scores (52 % correct), while the *F*0c carrier produced the highest tone recognition scores (95 % correct). The NB carrier produced moderate tone recognition scores (62 % correct). A two-way ANOVA showed that Chinese tone recognition was significantly affected by the carrier type [$F(2,30)=240.61$, $p<0.001$], but not by the number of channels [$F(1,30)=0.02$, $p=0.88$].

Figure 2.D shows Chinese sentence recognition scores (recorded as the percent of key words recognized in a set of 20 sentences). Sentence recognition significantly improved for all carriers as the number of channels was increased. The *F*0c carrier produced the highest sentence recognition scores (33 % correct with 2 channels; 81 % correct with 4 channels). The FR carrier produced the lowest sentence recognition scores (16 % correct with 2 channels; 46 % correct with 4 channels). Similar to tone recognition, the NB carrier produced moderate sentence recognition scores (27 % correct with 2 channels; 70 % correct with 4 channels). A two-way ANOVA revealed significant effects for both the number of channels [$F(1,30)=133.14$, $p<0.001$] and the carrier type [$F(2,30)=19.96$, $p<0.001$].

4. DISCUSSION

As the number of channels was increased, the representation of the spectral envelope and its transitions was enhanced. This spectral enhancement provided better Chinese vowel and consonant recognition. However, even with 4 channels, the spectral resolution did not provide additional pitch and periodicity information for Chinese tone recognition. Therefore, tone recognition was not affected by the number of spectral channels.

The results obtained with the experimental carriers simulate the effects of pitch and periodicity cues on Chinese speech recognition by CI users. Chinese vowel recognition was not affected by the carrier type, indicating that different amounts of pitch and periodicity information did not affect perception of the spectral envelope. In contrast, Chinese tone recognition primarily depends on the perception of pitch changes. The FR carrier, which provided a flat tone perception, produced the least amount of tone recognition. The NB carrier produced moderate tone recognition scores, most likely due to the weak pitch

information found in the 50 Hz low-pass temporal envelopes [3]. The *F*0c carrier produced almost perfect tone recognition, because the pitch and periodicity information was directly encoded in the carrier. When compared to English consonant recognition [2], there was a significant increase in Chinese consonant recognition when binary carriers (*F*0c and FR) were used instead of the NB carrier.

Chinese sentence recognition is comprised of phoneme recognition, tone recognition, and comprehension of sentence contexts. When the number of channels was increased from 2 to 4, Chinese sentence recognition improved greatly, largely because of improved phoneme recognition. By changing from the NB to *F*0c carrier, the increased pitch and periodicity information produced a further improvement in Chinese sentence recognition with 4 channels. The FR carrier produced the poorest tone recognition and therefore contributed little to Chinese sentence recognition, for which tone recognition is important. The relatively weak pitch cues found in the NB carrier produced moderate tone recognition, which was helpful in producing moderate levels of Chinese sentence recognition. These varying patterns in Chinese sentence recognition were not found for English sentence recognition [2], indicating the special importance of pitch and periodicity information in tonal languages such as Chinese.

The relative contributions of vowel, consonant, and tone recognition to Chinese sentence recognition may be quantified by using a modified Boothroyd power-function model [4], as shown in equation 1:

$$Ps = 1 - (1 - Pv^{Wv} Pc^{Wc} Pt^{Wt})^{K} \qquad (1)$$

where *Ps*, *Pv*, *Pc*, and *Pt* represent Chinese sentence, vowel, consonant, and tone recognition scores respectively, *Wv*, *Wc*, and *Wt* are the weights of vowel, consonant, and tone recognition for isolated Chinese syllable recognition, and *K* is a variable that describes the context effects (the number of isolated syllable recognition errors that cause sentence recognition errors).

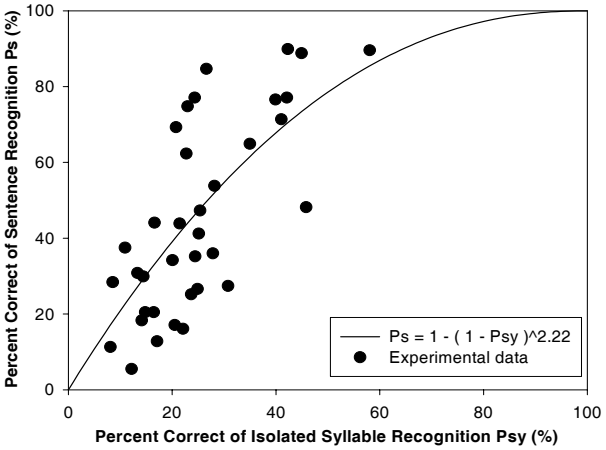

Figure 3: The relationship between Chinese isolated syllable recognition and word-in-sentence recognition.

Figure 3 shows the best fitting power-function (the solid line) relating isolated syllable recognition scores to word-in-sentence recognition scores (best fitting $r = 0.73$; parameters: $Wt = 0.76$, $Wv = 1.00$, $Wc = 0.77$, $K = 2.22$); the symbols represent experimental data. The similar weights of Chinese tone, vowel, and consonant recognition indicate their similar importance in Chinese speech recognition.

5. CONCLUSION

The results of the present study suggest that delivering more pitch and periodicity information to CI users may enhance patients' Chinese speech perception. Encoding pitch information in the *F*0c carrier provides the best tone and sentence recognition when limited spectral resolution is available. However, the *F*0c carrier is not suitable for implementation in CI devices because of its relatively low stimulation rates; furthermore, CI users have difficulty tracking pitch changes produced by variable stimulation rates. An effective pitch coding strategy in CI applications is necessary for Chinese-speaking CI users.

6. REFERENCE

[1] L. Sagart, "Tone production in modern standard Chinese: An electromyographic investigation," Cahiers de Linguistique, Asie Orientale, Paris, 205-221, 1986.

[2] A. Faulkner, S. Rosen, and C. Smith, "Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants," J. Acoust. Soc. Am. 108, 1877-1887, 2000.

[3] Q.-J. Fu, F.-G. Zeng, R. V. Shannon, and S. D. Soli, "Importance of tonal envelope cues in Chinese speech recognition," J. Acoust. Soc. Am. 104, 505-510, 1998.

[4] A. Boothroyd, and S. Nittrouer, "Mathematical treatment of context effects in phoneme and word recognition," J. Acoust. Soc. Am. 84, 101-114, 1988.

[5] R.-H. Wang, "The standard Chinese database," University of Science and Technology of China, internal materials, 1993.

[6] S. D. Soli, "Hearing in Noise Test for Mandarin Chinese," 2003.

[7] B. S. Wilson, C. C. Finley, D. T. Lawson, R. D. Wolford, D. K. Eddington, and W. M. Rabinowitz, "Better speech recognition with cochlear implants," Nature (London) 352, 236-238, 1991.

[8] D. D. Greenwood, "A cochlear frequency-position function for several species -- 29 years later," J. Acoust. Soc. Am. 87, 2592-2605, 1990.

[9] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," IEEE Trans. Audio Electroacoust, Vol. AU-20, No. 5, 367-377, 1972.

[10] G. Miller, and P. Nicely, "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. 27, 338-352, 1955.