International Congress Series 1273 (2004) 23–26

www.ics-elsevier.com

Temporal fine structure: the missing component in speech processing algorithms

G.S. Stickney[a,*], K. Nie[a], Y.-Y. Kong[b], H. Chen[c], F.-G. Zeng[a,b,c]

[a]*Department of Otolaryngology, University of California, Irvine, United States*
[b]*Department of Cognitive Sciences, University of California, Irvine, United States*
[c]*Department of Biomedical Engineering, University of California, Irvine, United States*

Abstract. A speech processing algorithm, which encodes the temporal fine structure of sound by extracting slowly varying frequency modulations, was shown to improve speech perception in noise, talker identification, melody recognition, and tonal language recognition.

Keywords: Cochlear implant; Temporal fine structure; Speech processing

1. Introduction

Multichannel cochlear implants (CIs) use multiple electrodes distributed along the cochlea to take advantage of the tonotopic organization of auditory nerve fibers. This coding allows most implant users to adequately perceive approximately 80% of the words for sentences presented in quiet. However, performance in noisy backgrounds and the sound quality of music presented through the device are severely limited.

A novel speech processing algorithm that may extend the capabilities of CIs is proposed. This algorithm utilizes both place and temporal coding to transmit frequency information. It is well known that varying the rate of stimulation changes the pitch: low stimulation rates yield the perception of a low-frequency sound and vice versa. The current study utilizes temporal coding to extract and transmit temporal fine structure. Since temporal fine structure conveys information about formant patterns, formant transitions,

* Corresponding author. Tel.: +1 949 824 3927; fax: +1 949 824 5907.
E-mail address: stickney@uci.edu (G.S. Stickney).

0531-5131/
doi:10.1016/j.ics.2004.09.017

and pitch, it can be important for speech recognition in noise, talker identification, melody recognition, and tonal language recognition.

The speech processing algorithm was based on Hilbert's definition of temporal fine structure [1]. According to this definition, sound can be decomposed into two components: the envelope and the temporal fine structure. The envelope can be conveyed by amplitude modulation (AM) over time, whereas temporal fine structure can be conveyed by frequency modulation (FM) over time [2].

2. Materials and methods

2.1. Subjects

Twenty-four normal-hearing (NH) and 5 CI subjects participated in the sentence recognition task, 5 NH and 10 CI subjects participated in the vowel and speaker identification task, 5 native NH Mandarin-speaking subjects participated in the tonal language recognition task, and 6 NH and 6 CI subjects participated in the melody recognition task. CI subjects had at least 6 months of experience with their device.

2.2. Processing

NH subjects were presented with an implant simulation consisting of AM information only or possessing both AM+FM information [3]. CI users were only presented with unprocessed stimuli. A diagram of the algorithm is shown in Fig. 1 [3]. The original stimulus was divided into 1–32 narrowbands and subjected to separate AM and FM extraction pathways. Along the AM processing pathway, the envelope was extracted through full-wave rectification and low-pass filtering with a 500-Hz cutoff. The FM, which represents the temporal fine structure, was derived by first removing the center frequency using phase orthogonal demodulators, followed by two low-pass filters. One low-pass filter set the FM bandwidth to 500 Hz (or the critical bandwidth, whichever was smaller) and the second filter set the FM cutoff to a slowly varying rate of 400 Hz. A slowly varying FM is critical since most implant listeners cannot track FM greater than 500 Hz [4]. The last step recovers the center frequency and uses it to recombine the

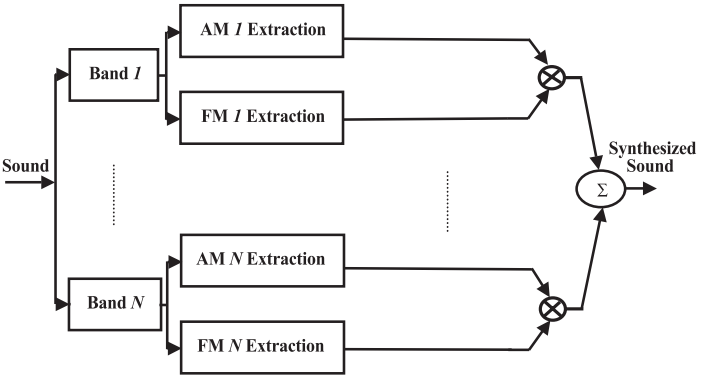

Fig. 1. The novel speech processing algorithm (called frequency amplitude modulation encoding, or FAME).

narrowbands from each extraction pathway. For the AM carrier, the FM bandwidth was set to 0 Hz, or a noise carrier was used.

2.3. Stimuli

Sixty sentences were presented in quiet or with a second competing sentence. Vowel identification consisted of 12 /hVd/ vowels spoken by three men, three women, two boys, and two girls. A subset of six vowels (three for a practice session and a different set of three for the test session) was used for speaker identification. Mandarin words were spoken in quiet, or with speech-shaped noise. In the melody identification task, 12 familiar melodies were presented with the rhythmic cue removed (accomplished by setting the duration of the notes to 350 ms with a silent interval of 150 ms between each note).

2.4. Procedures

In the sentence recognition experiment, subjects responded by typing in the words of the sentence. In all other tasks, subjects pressed a response button on the computer monitor. Practice sessions were given prior to testing. Responses were scored in terms of percent correct. Subjects were seated in a double-walled sound-attenuated chamber. Stimuli were presented to the NH listeners monaurally through a Sennheiser headphone at a level of 65 dB SPL. CI subjects were presented with sentence and music stimuli using a direct connection at a comfortable listening level and through a TANNOY Reveal speaker for the vowel and speaker identification task (0° azimuth; 65 dB SPL).

3. Results

All experiments showed a significant improvement in performance when FM information was added to the simulation. Fig. 2a and b shows results for sentences processed into four-channel (Fig. 2a) and eight-channel (Fig. 2b) conditions as a function of the signal-to-noise ratio (SNR). The benefit provided by FM was most evident for the

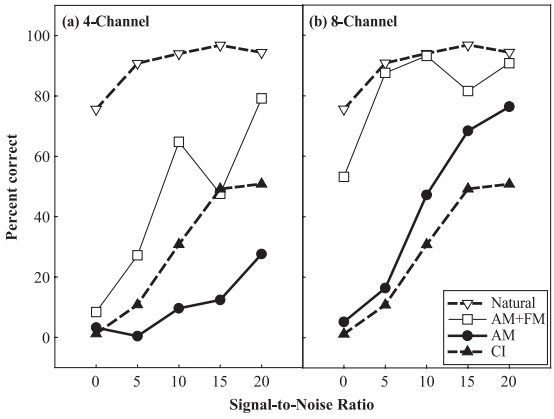

Fig. 2. Sentence recognition for NH listeners presented with an implant simulation consisting of AM information only (filled circles) or AM+FM information (unfilled squares). Results for natural speech (unfilled triangles) and from CI users (filled triangles) are included in each panel for comparison.

2-, 4-, 8-, and 16-channel conditions. Results for CI listeners were within the range of performance obtained by NH listeners presented with 8–16 AM channels in quiet and with four to eight AM channels at a 10-dB SNR. This range is where the additional FM cue provided the greatest benefit with the simulation. FM information also improved performance in the speaker identification and vowel recognition task, although to a much greater degree in the former [5]. The performance of CI users mirrored that of the NH listeners presented with four to eight AM channels in the vowel recognition task, and with one AM channel in the more difficult speaker identification task. As with the previous two experiments, the results for Mandarin tones demonstrated a significant benefit with the additional FM cue, and the benefit was most pronounced in the noise condition [5]. Finally, with the rhythmic cue removed in the melody recognition task, 32 bands were needed to reach maximum performance when only AM information was provided. However, with the addition of FM, only four bands were needed to reach maximum performance. CI performance for melody recognition similar to NH listeners presented with 8–16 AM-only channels.

4. Discussion

These results demonstrate that the addition of temporal fine structure to current speech processing algorithms has the potential to improve CI performance. Temporal fine structure complements the envelope information currently provided by CIs, and, as demonstrated here, can be useful for speech recognition in noise, talker identification, music, and tonal language perception.

FM is a slowly changing version of temporal fine structure that can be perceived by CI users. Since most CI users can only follow frequency changes coded temporally up to at most 500 Hz, the FM cutoff should be limited by the upper temporal frequency discrimination boundary. To code temporal fine structure in speech processing strategies such as continuous interleaved sampling (CIS), the interval between successive pulses in the pulse train carrier would be varied according to the FM, while the intensity of the signal would be conveyed with AM. Psychophysical FM tasks in CI users are currently underway to assess the feasibility of this novel speech processing algorithm [6].

References

[1] D. Hilbert, Grundzuge einer allgemeinen, Theorie der linearen Integralgleichungen, Teubner, Leipzig, 1912.
[2] P. Loughin, B. Tacer, On the amplitude-and-frequency modulation decomposition of signals, J. Acoust. Soc. Am. 100 (1996) 1594–1601.
[3] K. Nie, G. Stickney, F.G. Zeng, Encoding frequency modulation to improve cochlear implant performance in noise, IEEE Trans. Biomed. Eng. (in press).
[4] F.G. Zeng, Temporal pitch in electric hearing, Hear. Res. 174 (2002) 101–106.
[5] Y. Kong, M. Vongphoe, F.G. Zeng, Independent contributions of amplitude modulation and frequency modulation to auditory perception II. Melody, tone and speaker identification, Abstract of the 26th ARO midwinter meeting, 2003, pp. 213–214.
[6] H. Chen, F.G. Zeng, Frequency modulation detection in cochlear implant subjects, J. Acoust. Soc. Am. (in press).