Demonstration of a Novel Speech-Coding Method for Single-Channel Cochlear Stimulation

Yuta TAMAI*, Shizuko HIRYU*, and Kohta I. KOBAYASI*

(Received October 20, 2017)

Cochlear implants are widely used to restore hearing, but implanting an electrode into the cochlea is a major surgical intervention. Infrared lasers generate compound action potentials from neurons with heat-sensitive channels and stimulate nerves without contacting tissues; therefore, an infrared laser-based hearing aid can produce auditory perception without requiring electrode insertion into the cochlea. The purpose of this study was to develop a speech-coding scheme using infrared laser stimulation to convey intelligible speech. Because the laser stimulation has a limited capacity to stimulate auditory neurons differentially, it evokes action potentials from the entire cochlear nerve simultaneously. Thus, the stimulation may be perceived as a clicking sound. We therefore created a click-modulated speech sound to simulate infrared laser stimulation. The sound was a repetitive click with a pitch (repetition rate) and amplitude similar to the formant frequency and amplitude envelope transition of an original speech sound, respectively. Five Japanese native speakers with normal hearing participated in the experiment. First, the subjects listened to the click-modulated speech sounds and answered questions about what they perceived, selecting from four choices; second, they listened to the sound and wrote down what they perceived, using the Roman alphabet. The percentages of correct answers on both the alternative-choice and dictation tests were significantly higher than the rates obtained by chance. Our data suggests that click-modulated speech sounds were at least partially intelligible, and that a hearing aid with an infrared laser could restore speech perception in hearing-impaired individuals.

Key words : infrared laser, noninvasive stimulating system, click-modulated speech sound

1. Introduction

Cochlear implants are widely used to compensate for hearing loss by transforming sound information into neural electrical activity. However, a cochlear implant requires invasive surgery because an electrode must be inserted into the cochlea to stimulate the cochlear nerve. This procedure has the risk of further hearing loss.

A previous study demonstrated that action potentials are evoked by irradiating neurons with an infrared laser[1]. Infrared laser stimulation is gaining significant attention as a potential substitute for electrical stimulation because a laser can stimulate nerves without contacting neurons[1]. Izzo and co-workers (2006, 2007, 2008) observed compound action potentials by irradiating the cochlear nerve with an infrared laser[2-4]. That group also investigated auditory perception generated by an infrared laser.

The purpose of this study was to develop a speech-coding scheme for an infrared laser-based hearing aid. Infrared laser stimulation can create auditory perception similar to that evoked by single-channel cochlear stimulation because laser irradiation evokes almost all cochlear nerve responses simultaneously[5]. A pioneering study by Fourcin and Rosen (1979) showed that a single electrode placed at the round window was able to produce various acoustic features of speech such as intonation and voiced–voiceless information[6]. Another study revealed

*Neuroethology and Bioengineering Laboratory, Department of Biomedical Information, Doshisha University, Kyo-tanabe, Kyoto, 610-0321
Telephone: +81-774-65-6439, E-mail: dmq1041@mail4.doshisha.ac.jp

that extra-cochlear single-channel implants improved lip-reading ability[7]. Despite these early successes, an extra-cochlear implant has not been fully implemented clinically. Because an extra-cochlear single-channel system stimulates all cochlear nerve fibers simultaneously, the system cannot replicate the fine frequency structure, and is regarded as less capable of restoring speech perception than the multi-channel system. As the intelligibility of an extra-cochlear single-channel system is not sufficient to produce speech perception clearly, few hearing-impaired individuals wear single-channel cochlear implants.

In this study, we investigated the speech-coding schemes of an infrared laser hearing aid. Because it is difficult to stimulate the cochlear nerve differentially using an infrared laser (i.e., single-channel stimulation), we assumed that the auditory stimulation is perceived as a clicking sound. We synthesized a click-modulated speech sound (CMS) as a simulated sound in a click train, with a pitch (repetition rate) and amplitude similar to the formant center frequency (F1 and F2) and amplitude envelope transition of an original speech sound.

2. Materials and Methods

2.1 Subjects

Five native Japanese speakers (22–27 years old) participated in the experiment. None of the subjects had listened to CMSs before participating in the study, and all passed a hearing screening test at a 25 dB hearing level at frequencies of 0.5, 1, 2, and 4 kHz.

2.2 Stimuli

Click-modulated speech sound

We synthesized a CMS that simulates the perception evoked by irradiating a cochlear bundle with an infrared laser. The sound is a click train with a pitch (repetition rate) similar to the formant center frequency of the first and second formants of an original speech sound, and with an amplitude envelope simulating the original envelope. Specifically, formant frequencies were extracted from the original sounds by linear predictive coding and fast Fourier transforms at a 48-kHz sampling rate and 1024-point fast Fourier transform length. Linear predictive coding was calculated every 15 ms over 30-ms Hamming windowed segments. The amplitude envelope was extracted using a low-pass filter (cutoff frequency = 46 Hz) after half-wave rectification (Fig. 1). All signal processing was performed using Matlab (MathWorks). An example of an original sound and a click-modulated speech sound was shown in Fig.2. The original speech sounds were Japanese four-mora words spoken by a female speaker. All sounds were obtained from a publicly available data set of familiarity-controlled word lists (FW03)[8]. We randomly selected 50 words from the high-familiarity list (word-familiarity rank of 5.5–7.0)[8]. Examples of the stimulus words were “a.ma.gu.mo,” “ki.ta.ka.ze,” and “pa.chi.n.ko.”

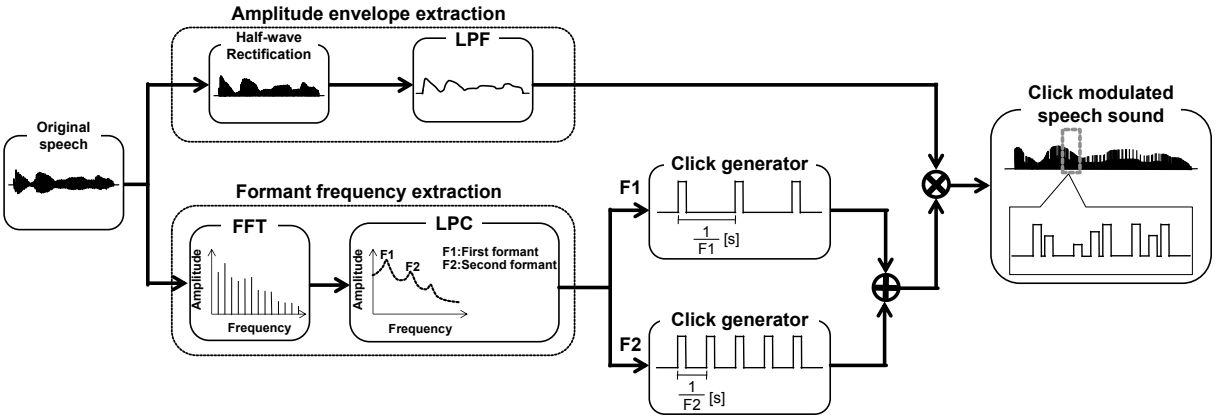

Fig. 1. Coding process of a click-modulated speech sound (CMS). Schematic diagram depicting the process of analyzing the speech signal and synthesizing the CMS. See text for details.

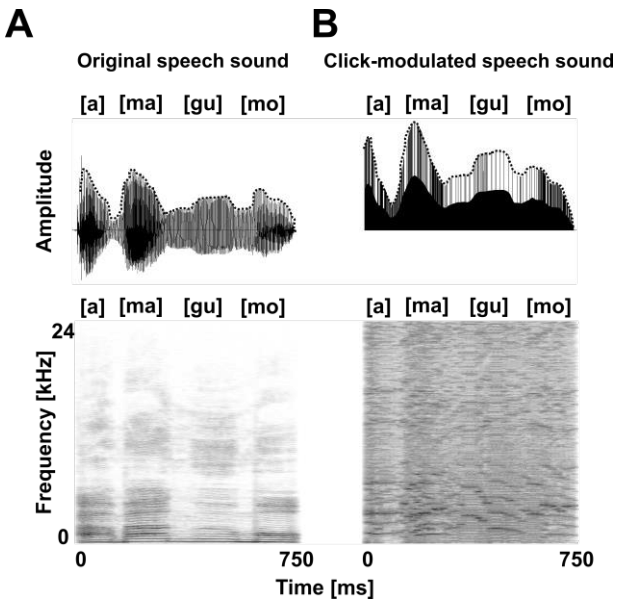

Fig. 2. An example of the stimulus. The upper graphs show the waveform and amplitude envelope (dotted line), and the lower figures are spectrograms. (A) The original speech sound used: the Japanese words "[a][ma][gu][mo]." (B) Click-modulated speech sound synthesized from the original sound in (A).

2.3 Experimental environment

All experiments were conducted in a sound-proofed room. The stimulus was presented to the subject through headphones (STAX Lambda Nova, STAX Industries) using a D/A converter (Octa-capture, Roland). The sound pressure level of all stimuli was measured using a microphone (ER-7C Series B, Etymotic Research) and calculated at sound pressure level of 60 dB.

2.4 Experimental procedure

Two experiments were conducted. Subjects first participated in an alternative-choice task and then proceeded to a dictation task. In the alternative-choice task, four choices written in Japanese were presented on a printed sheet before the subjects listened to each CMS. The subjects were instructed to answer questions from among four choices regarding the sound they perceived within 5 s after the stimulus was presented. In the dictation task, the subjects listened to the simulated sounds and wrote down their perceptions on response sheets using Roman letters. Fifty trials (= 50 words) were conducted in both the alternative-choice and dictation tasks.

Table 1. Examples of the Japanese stimulus words. Fifty Japanese words were randomly selected from a familiarity-controlled database of four-mora words (FW03).

Word	Mora	Articulation
雨雲（アマグモ）	[a][ma][gu][mo]	a.ma.gu.mo
今風（イマフウ）	[i][ma][fu][u]	i.ma.fu.:
内側（ウチガワ）	[u][chi][ga][wa]	u.chi.ga.wa
押出（オシダシ）	[o][shi][da][shi]	o.shi.da.shi
親元（オヤモト）	[o][ya][mo][to]	o.ya.mo.to
警察（ケイサツ）	[ke][i][sa][tu]	ke.:.sa.tu

3. Results

Figure 3A shows the results of the alternative-choice task. The highest percentage of correct answers among all subjects was 88%, and the lowest was 44%. The average percentage of correct answers among the five subjects was 74%, which was significantly higher than the chance level (25%; one-sample t-test, $p < 0.001$). Figure 3B shows the rate of correctly perceived vowels, consonants, and morae. In the rate of correctly perceived vowels, the highest percentage of correct answers among all subjects was 63%, and the lowest was 37%. The average percentage of correct answers among the five subjects was 49%. A crude estimate of obtaining the correct answers by chance in this task was 20%, since subjects choose one vowel from only five ('a', 'i', 'u', 'e', and 'o'). The average percentage of correct answers was significantly higher than the chance level (one-sample t-test, $p < 0.001$). In the rate of correctly perceived consonants, the highest percentage of correct answers among all subjects was 36%, and the lowest was 11%. The average percentage of correct answers among the five subjects was 23%. A crude estimate of obtaining the correct answers by chance is 6%, as there are 16 Japanese consonants ('y', 'k', 'g', 's', 'z', 't', 'd', 'n', 'h', 'p', 'b', 'm', 'r', 'w', 'N', and 'Q'). The average percentage of correct answers was significantly higher than the chance level (one-sample t-test, $p < 0.01$). In the rate of

correctly perceived morae, the average percentage of correct answers was 19%, ranging from 11% to 28%.

4. Discussion

Our data showed that the subjects were able to understand the contents of the click-vocoded speech sounds at least to some extent (Fig. 3). As demonstrated by many previous studies, perceptions of syllables are strongly related to formant frequency[9]. For vowels, the first and second formants are almost sufficient to distinguish different words; even for consonants, the involvement of the third or higher formant in perception is relatively limited[10]. Remez et al. (1981) developed a synthetic sound composed of several time-varying sine waves, each of which replicated the frequency and amplitude patterns of the formants[11]. This sound is called sine-wave speech, and its intelligibility has been well described[11,12]. The CMS used in this experiment is akin to sine-wave speech composed of two sinusoids corresponding to the first and second formants. The sine-wave speech was partially comprehensible, and the syllable transcription rate was approximately 36%[11]. This value is comparable with our results (Fig. 3B), suggesting that the CMS may convey the same acoustic key (i.e., formant frequency) as that of sine-wave speech.

As shown in Fig. 3A, the subjects were able to perceive the CMS relatively easily if the correct answer was provided as an option. The relatively high performance on the alternative-choice task compared with the dictation task (Fig. 3B) may be partly due to top-down cognitive processing. Speech perception is complex by nature, involving both bottom-up and top-down information processes. Bottom-up processing includes the physical acoustic characteristics such as time-varying formant information and the amplitude envelope. Top-down processing involves perceptual cognitive information such as lexical segmentation and perceptual learning[13]. Several studies have revealed the effects of these top-down processes on the comprehension of distorted speech sounds[12,13]. In one of the most extreme cases, sine-wave speech was totally incomprehensible if the subject did not hear the sound as speech, but it became almost completely comprehensible once the subject was aware that it was speech[11]. In the alternative-choice task, subjects were given the possible answer choices before they listened to the CMS, and that information may have assisted perception through top-down processes.

For practical purposes, the intelligibility of the CMS needs to be improved. We can pursue several possibilities. The harmonics of the first formant of the CMS overlaps acoustically with the second formant in most cases (Fig. 2B); this overlap could mask the

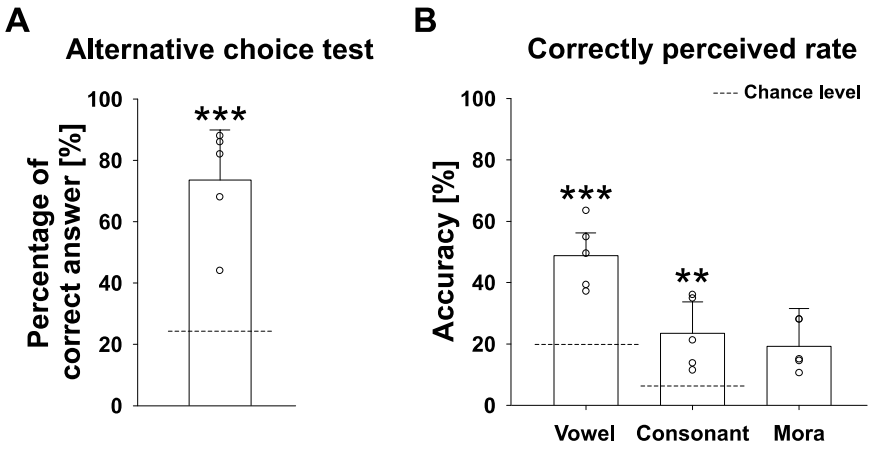

Fig. 3. Intelligibility of a click-modulated speech sound. Error bars represent the standard deviations of the mean. Average scores were compared with the chance level (CL) using a one-sample t-test (** <0.01, *** <0.001). (A) Results of the alternative-choice task. Vertical axis shows the percentage of correct answers. The CL is 25%. (B) The rates of correctly perceived vowels, consonants, and morae in the dictation task. The CL of the rate of correctly perceived vowels is 20%. The CL of the rate of correctly perceived consonants is 6%.

perception of the second formant. Therefore, amplifying the second formant may improve intelligibly by decreasing the masking. In addition to modifying the acoustical parameters, combining audio and visual information may facilitate perception. Of the many studies on multimodal speech perception, a pioneering study by Sumby and Pollack (1954) demonstrated that speech perception in noisy environments was improved by as much as +15 dB when the listener could see the speaker's face[14]. Visual information such as the dynamics of an articulating face allows individuals with hearing disabilities to identify the phonetic and lexical information of speech sounds[15]. Comprehension of sine-wave speech was significantly improved when articulatory location (i.e., movement of the mouth) was provided along with the sound stimulus[12]. The intelligibility of a CMS, therefore, may improve if the subject can see the movements of the articulator (tongue and lips). Because in most real-life situations the listener is able to see the speaker's face, we could expect better performance comprehending a CMS (infrared cochlear stimulation) in everyday life compared with our experiment. The efficacy of a CMS in multimodal situations needs to be investigated in future studies.

5. Conclusion

In this experiment, we quantified the intelligibility of a CMS. The sound was designed to simulate the sensation evoked by an infrared laser-based hearing aid. This speech coding scheme is in line with previous research on distorted speech sounds, such as noise-vocoded and sine-wave speech sounds. All participants were able to comprehend the contents of the CMS, at least partially. The intelligibility of the sounds was comparable with that of conventional sine-wave speech, suggesting that the participants mainly used the formant frequency as their cue to comprehend the sound. Overall, our results demonstrate that even severely distorted speech sounds are intelligible if the sound retains time-varying formant information and temporal amplitude envelopes, and indicates that an infrared laser hearing aid with CMS is a potential non- or less-invasive alternative to conventional cochlear implants for restoring speech perception.

References

1) J. Wells *et al.*, "Optical Stimulation of Neural Tissue in Vivo", *Opt. Lett.*, **30** [5], 504–506 (2005).
2) A. D. Izzo, C. P. Richter, E. D. Jansen, and J. T. Walsh, "Laser Stimulation of the Auditory Nerve", *Lasers Surg. Med.*, **38** [8], 745–753 (2006).
3) A. D. Izzo *et al.*, "Optical Parameter Variability in Laser Nerve Stimulation: A Study of Pulse Duration, Repetition Rate, and Wavelength", *IEEE Trans. Biomed. Eng.*, **54** [6], 1108–1114 (2007).
4) A. D. Izzo *et al.*, "Laser Stimulation of Auditory Neurons: Effect of Shorter Pulse Duration and Penetration Depth", *Biophys. J.*, **94** [8], 3159–3166 (2008).
5) Y. Tamai, Y. Shinpo, K. Horinouchi, S. Hiryu, and K. I. Kobayasi, "Infrared Neural Stimulation Evokes Auditory Brain Stem Responses following the Acoustic Feature of Speech Sounds", *The Harris Science Review of Doshisha University*, **57** [4], 254–261 (2017).
6) A. Fourcin and S. Rosen, "External Electrical Stimulation of the Cochlea: Clinical, Psychophysical, Speech-Perceptual and Histological Findings", *Br. J. Audiol.*, **13** [3], 85–107 (1979).
7) S. Rosen and V. Ball, "Speech Perception with the Vienna Extra-Cochlear Single-Channel Implant: a Comparison of Two Approaches to Speech Coding", *Br. J. Audiol.*, **20** [1], 61–83 (1986).
8) S. Sakamoto, Y. Suzuki, S. Amano, and K. Ozawa, "New Lists for Word Intelligibility Test Based on Word Familiarity and Phonetic Balance", *J. Acoust. Soc. Jpn*, 842–849 (1998).
9) J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic Characteristics of American English Vowels", *J. Acoust. Soc. Am.*, **97** [5 Pt 1], 3099–3111 (1995).
10) K. Miyawaki, W. Strange, R. Verbrugge, A. M. Liberman, J. J. Jenkins, and O. Fujimura, "An Effect of Linguistic Experience: The Discrimination of [r] and [1] by Native Speakers of Japanese and English", *Percept. Psychophys.*, **18** [5], 331–340 (1975).
11) R. E. Remez, P. E. Rubin, D. B. Pisoni, and T. D. Carrell, "Speech Perception without Traditional Cues", *Science*, **212** [22]. 947–950 (1981).
12) R. E. Remez, J. M. Fellowes, D. B. Pisoni, W. D. Goh, and

P. E. Rubin, “Multimodal Perceptual Organization of Speech: Evidence from Tone Analogs of Spoken Utterances”, *Speech Commun.*, **26** [1], 65–73 (1998).

13) M. H. Davis and I. S. Johnsrude, “Hearing Speech Sounds: Top-Down Influences on the Interface between Audition and Speech Perception”, *Hear. Res.*, **229** [1–2], 132–147 (2007).

14) W. H. Sumby and I. Pollack, “Visual Contribution to Speech Intelligibility in Noise”, *J. Acoust. Soc. Am.*, **26** [2], 212 (1954).