Scientific Research Publishing

Open Journal of Acoustics, 2018, 8, 52-60
http://www.scirp.org/journal/oja
ISSN Online: 2162-5794
ISSN Print: 2162-5786

Speech Encoding Scheme for the Extra-Cochlear Pulsed Electrical Stimulation System

Yuta Tamai, Kazuyuki Matsumoto, Shizuko Hiryu, Kohta I. Kobayasi

Life and Medical Science, Doshisha University, Kyoto, Japan
Email: kkobayas@mail.doshisha.ac.jp

How to cite this paper: Tamai, Y., Matsumoto, K., Hiryu, S. and Kobayasi, K.I. (2018) Speech Encoding Scheme for the Extra-Cochlear Pulsed Electrical Stimulation System. *Open Journal of Acoustics*, **8**, 52-60.
https://doi.org/10.4236/oja.2018.83005

Received: June 10, 2018
Accepted: September 4, 2018
Published: September 7, 2018

Abstract

An extra-cochlear stimulation system has been investigated as a less invasive alternative to conventional cochlear implant; however, the system is used primarily as a speech-reading aid. The purpose of this study was to develop a speech encoding scheme for the extra-cochlear stimulation system to convey intelligible speech. A click-modulated speech sound (CMS) was created as a simulation of the extra-cochlear stimulation system. The CMS is a repetitive click with a repetition rate similar to the formant frequency transition of an original sound. Seven native Japanese speakers with normal hearing participated in the experiment. After listening to the CMS, synthesized from low familiarity Japanese words, the subjects reported their perceptions. The results showed that the rates of correctly identified vowels and consonants were significantly higher than those of the control stimulus, suggesting that the CMS can generate at least partially intelligible vowel and consonant perceptions. In all, the speech encoding scheme could be applied to the extra-cochlear stimulation system to restore speech perception.

Keywords

Auditory Prosthesis, Cochlear Implant, Noninvasive Stimulation System, Speech Perception

1. Introduction

Cochlear implants are widely used to compensate for sensorineural hearing loss. These devices restore hearing in otherwise deaf individuals. In the cochlea, low-frequency sounds activate neurons in the apex of the cochlea, and high-frequency sounds stimulate the basal portion of the cochlea; this organization is known as tonotopicity. A multi-channel electrode is inserted into the cochlea to restore the tonotopic responses of a normal acoustically stimulated

cochlea [1]. By restoring tonotopicity, the cochlear implant can produce detailed frequency information, and many cochlear implant users perceive electrical stimulation produced by the system as speech sounds.

One of the greatest drawbacks of the cochlear implant is that it requires major surgical intervention. Since the early stages of its development, the extra-cochlear implant has been considered a potential alternative to a multi-channel implant. A pioneering study by Fourcin and colleagues (1979) showed that a single electrode placed in the round window produces various acoustic features of speech, such as intonation and voiced-voiceless information [2]. Other studies revealed that extra-cochlear single-channel implants improved lip-reading ability [3]. Despite these early successes, the extra-cochlear implant has not been fully implemented clinically. Because the extra-cochlear single-channel system stimulates all cochlear nerve fibers simultaneously, it cannot replicate fine frequency structure, and thus this system is less capable of restoring speech perception compared with the multi-channel system. Thus far, extra-cochlear implants have been primarily used as a speech-reading aid.

In this study, we attempted to improve the speech encoding schemes of a single-channel stimulation system. Because very few individuals have extra-cochlear implants, the intelligibility of simulated sounds was tested in subjects with normal hearing. As mentioned earlier, using a single-channel stimulation system, it is difficult to stimulate the cochlear nerve differentially to replicate tonotopicity; instead, the system stimulates the entire cochlear nerve simultaneously. We therefore assumed that single-pulse electrical stimulation by an extra-cochlear electrode creates a perception similar to a clicking sound, and that continuous electrical stimulation is perceived as a series of clicks.

Here, we synthesized a click-modulated speech sound (CMS) as a simulated sound of a single-channel stimulation system. The sound was a click train with a pitch (repetition rate) similar to the formant center frequency of an original speech sound. As a first step to show feasibility of the speech encoding scheme, this research tested only normal hearing subjects as has been done in several previous researches [4] [5], and we focused on how first and second formant frequency, well known minimum requirements for stable speech perception [6], contribute the intelligibility of the CMS.

The CMS is acoustically similar to a sine-wave speech sound (SWS) in a sense that both sounds replicate the formant frequency of an original sound [7], and several studies have demonstrated SWSs to be intelligible [8] [9] [10]. Thus, we reasonably expected the CMS to be intelligible as a speech sound, and that the speech encoding scheme will revitalize clinical use of the extra-cochlear stimulation system.

2. Materials and Methods

2.1. Subject

Seven native Japanese speakers (21 - 24 years old) participated in the experi-

ment. None of the subjects had listened to the CMSs prior to the study, and all passed a hearing screening using a threshold hearing level (HL) of 25 decibel (dB) HL at 0.5, 1, 2, and 4 kHz frequencies.

2.2. Stimuli

2.2.1. Click-Modulated Speech Sound

We synthesized a CMS, which is a click train with a repetition rate similar to the formant center frequency of an original speech sound. **Figure 1** depicts how to synthesize the CMS. First formant (F1) and second formant (F2) frequencies were extracted by a publicly available MATLAB-based script. The program was commonly used and described in several previous studies [11] [12] [13]. It implements liner predictive coding (LPC) analysis, and LPC was calculated every 2.7 ms over 5.4 ms Hanning windowed segments at 8 kHz sampling rate with an LPC order equal to 8. **Table 1** lists the signal processing parameters used in the program. These analytic parameters were identical to previous researches [11] [12] [13], and which allows us to compare our data with these results. After formant frequencies extraction, click train with a repetition rate similar to the formant frequencies were generated. These click trains were combined for creating CMS. We synthesized three types of CMSs: F1CMS, in which the click repetition rate temporally followed the first formant frequency of original speech; F1F2CMS, which was created by summing two click trains, each replicating the temporal structures of first and second formant frequencies (**Figure 1**); constant frequency click-modulated speech sound (CFCMS), which has a constant repetition rate similar to the average of first formant frequency. CFCMS was used as a control stimulus to evaluate the comprehension of each synthesized speech sound.

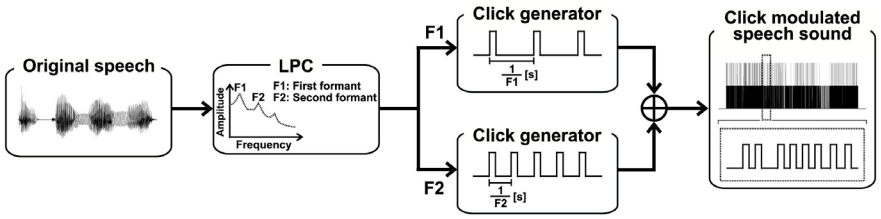

Figure 1. Process of encoding the click-modulated speech sound (CMS). Schematic diagram depicting the processes of analyzing the speech signal and synthesizing the CMS of every 2.7 ms segment. See text for details.

Table 1. The condition for creating click modulated speech sound (CMS).

Parameter	Variable
Analysis window	Hanning
Window length	5.4 ms
Shift length	2.7 ms
Pulse width	100 μs
LPC order	8

Original speech sounds were Japanese four-mora words voiced by a female speaker. All sounds were obtained from a publicly available data set of familiarity-controlled word lists (FW07) [14]. We randomly selected 24 words from the low familiarity list (word-familiarity rank of 1.0 - 2.5), and examples were shown in Table 2. Three kinds of CMSs were generated for each original sound (F1CMS, F1F2CMS, and CFCMS). Examples of an original sound and a CMS are shown in Figure 2.

2.2.2. Sine-Wave Speech Sound

The same words that produce CMSs were converted into SWSs. The SWS is a sound composed of various sine-waves that follow time-varying formant frequencies [7]. First and second formant frequencies were extracted from an

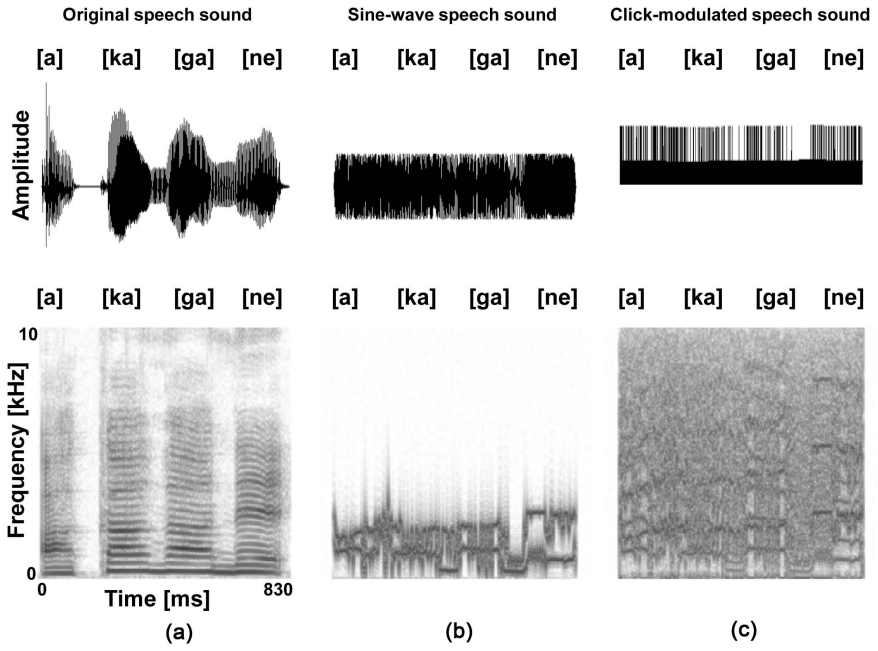

Figure 2. An example of the stimulus. Waveforms are presented in the upper figures and spectrograms in the lower figures. (a) original speech sound, (b) sine-wave speech sound, and (c) click-modulated speech sound of a Japanese word "[a] [ka] [ga] [ne]". (b) and (c) were synthesized from the original sound (a).

Table 2. Examples of Japanese stimulus words. Thirty-six Japanese words were randomly selected from a familiarity-controlled database of four-mora words (FW07).

Word	Mora	Meaning (English)
赤金 (アカガネ)	[a] [ka] [ga] [ne]	Gold-copper alloy
藁灰 (ワラバイ)	[wa] [ra] [ba] [i]	Straw ash
在方 (ザイカタ)	[za] [i] [ka] [ta]	Countryside
川淀 (カワヨド)	[ka] [wa] [yo] [do]	Stagnation point of a river
高殿 (タカドノ)	[ta] [ka] [do] [no]	High building
杣山 (ソマヤマ)	[so] [ma] [ya] [ma]	A wooded mountain

original sound using the same method as that for the CMS. Unlike most previous studies [7] [8] [9] [10], the time-varying amplitude was not replicated in this study, because the primary research focus was on perceptual contributions of formant trajectory created by click train and sinusoid sounds. As with the CMS, three types of stimuli were synthesized (F1SWS, F1F2SWS, and CFSWS). All SWS stimuli had 10 ms rise and fall times.

2.3. Experimental Procedure

All experiments were conducted in a soundproofed room. The stimulus was presented via headphones (STAX Lambda Nova, STAX Industries) with a digital-to-analog converter (Octa-capture, Roland). The sound pressure level of all stimuli was measured using a microphone (ER-7C Series B, Eatymotic Research) and calculated at 60 dB sound pressure level (SPL). All subjects were informed that they would be presented with 4-mora Japanese words as the stimulus and were instructed to write down their perceptions on response sheets using Roman letters within 10 s of stimulus presentation. Practice trials were conducted before the experiment, and subjects listened to six different stimuli (F1CMS, F1F2CMS, CFCMS, F1SWS, F1F2SWS, and CFSWS) three times without correct-answer feedback. In the practice trials, three different original sounds corresponded to F1, F1F2, CF conditions respectively, and these original sounds were converted into CMSs and SWSs. These stimuli were not used in the actual experiment. All subjects participated in four sessions, and 36 trials (=36 words) were conducted in each session.

2.4. Statistical Analysis

The mean (±standard deviation) rate of correct answers was calculated for each stimulus. A two-way analysis of variance (ANOVA) was performed using the stimulus type (CMS or SWS) and formant frequency condition (CF, F1, or F1F2) as independent variables, followed by pairwise comparisons using Tukey's honestly significant difference test (with a 5% level of significance). The Tukey's honestly significant difference test (Tukey's HSD) is one of a post hoc test. The test is commonly performed after an ANOVA for controlling the possibility of Type I errors while testing all pairwise difference [15]. All analyses were performed using a commercially available statistical program (SPSS, IBM, Armonk, NY, USA).

3. Results

Figure 3(a) shows the rates of correctly perceived vowels for the SWS and CMS; this rate is a reflection of the subject's perception of vowels, irrespective of consonants. The average rates of correct answers were 22% (CFSWS) and 23% (CFCMS) for the CF stimulus, 25% (F1SWS) and 23% (F1CMS) for the F1 stimulus, and 33% (F1F2SWS) and 30% (F1F2CMS) for the F1F2 stimulus. ANOVA revealed that the effect of stimulus type (CMS or SWS) was not

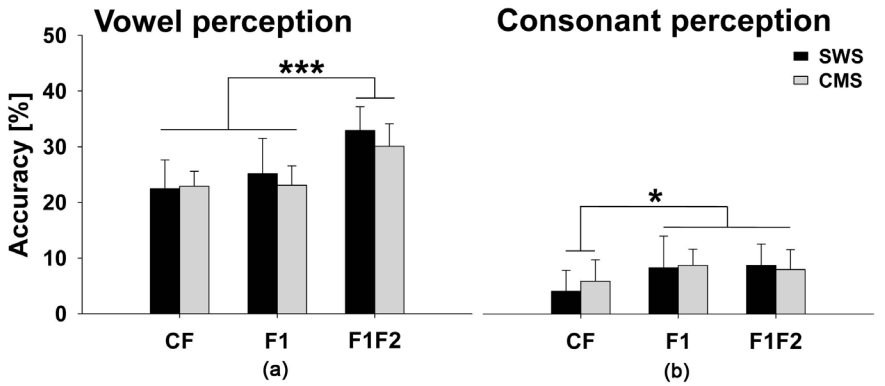

Figure 3. Intelligibility of sine-wave speech and click-modulated speech sounds. Error bars represent the standard deviation of the mean. Two-way ANOVA was performed to evaluate the effects of stimulus type and formant frequency condition. Regarding the intelligibility of vowel sounds, significant differences were observed between CF andF1F2 and between F1 and F1F2 (***$p < 0.001$). Regarding the intelligibility of consonants, CF and F1 were significantly different (*$p < 0.05$). (a) the rates of correctly perceived vowels; (b) the rates of correctly perceived consonants.

statistically significant, while the effect of formant frequency (CF, F1, or F1F2) was significant ($F_{2,30} = 18.16$, $p < 0.001$); multiple comparisons showed that the accuracy of the F1F2 condition was significantly higher than those of the F1 and CF conditions ($p < 0.001$).

Figure 3(b) shows the rates of correctly perceived consonants for the SWS and CMS; this rate is a reflection of the subject's perception of consonants, irrespective of vowels. The mean percentages of correct answers were 4.1% (CFSWS) and 5.9% (CFCMS) for the CF stimulus, 8.3% (F1SWS) and 8.7% (F1CMS) for the F1 stimulus, and 8.7% (F1F2SWS) and 8.0% (F1F2CMS) for the F1F2 stimulus. ANOVA did not reveal a significant effect of stimulus type (CMS or SWS), while the effect of formant frequency (CMS, F1, or F1F2) was significant ($F_{2,30} = 6.32$, $p < 0.01$); multiple comparisons showed that the accuracy of the CF condition was significantly lower than those of the F1 and F1F2 conditions ($p < 0.05$).

4. Discussion

Our data showed that subjects were able to perceive the CMS content at least to some extent (**Figure 3**). As many previous studies have demonstrated, perceptions of syllables are strongly related to formant frequency [6] [16]. Remez and colleagues (1984) developed distorted speech sounds combining several sine waves, each of which replicated time-varying formant frequencies (SWS), and demonstrated that the sound was partially comprehensible, with a syllable comprehension rate of approximately 36% [7]. In terms of information, the CMS we used in this experiment is similar to the SWS, because both sounds have the information provided by time-varying formant frequencies. **Figure 3(a)** and **Figure 3(b)** showed that the rates of correctly perceived vowels and consonants were not significantly different between the SWS and CMS, suggesting that the

CMS can be perceived as a speech sound similarly to the SWS.

As shown in **Figure 3(b)**, the CMS generated at least a partially intelligible perception of consonants. Several previous studies have reported F1 as a cue to categorize voiced and voiceless syllables and a place of articulation in stop consonants [17] [18] [19]. A significant improvement in F1CMS from CFCMS could reflect the information on voicing and place of articulation provided by F1. **Figure 3(b)** also showed that the difference in accuracy between the F1 and F1F2 conditions was not statistically significant ($p > 0.05$). To improve the comprehension of consonants, there are several straightforward improvements that we can apply. Implementing a higher fundamental frequency (*i.e.* F3) and/or the temporal envelope of each formant frequency would increase the comprehension of consonants.

In addition to modifying the acoustical parameters, combining audio and visual information can facilitate perception. Many studies on multimodal speech perception have been conducted; a pioneering study by Sumbyand (1954) demonstrated that speech perception in noisy environments is improved by as much as +15 dB when listeners were able to see the speaker's face [20]. 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 [21]. The comprehension of sine-wave speech was significantly improved when the articulatory location (*i.e.* movement of the mouth) was provided with the sound stimulus [8]. The intelligibility of the CMS, therefore, was improved if the subject could see the movements of the articulator. As subjects are able to see the speaker's face in most real-life situations, we expect better performance with the CMS in everyday use than in our experiment. However, because this research dealt only with normal hearing subject, further studies involving cochlear implantees is needed to evaluate the efficacy of our encoding scheme.

5. Conclusion

In this experiment, we quantified the intelligibility of the CMS. The sound was designed to simulate the sensation evoked by single-channel stimulation of the cochlear bundle. 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. In summary, our results demonstrate that single-channel stimulation with the CMS may be a non- or less-invasive alternative to conventional cochlear implants for restoring speech perception.

Acknowledgements

The authors thank Airi Ito for technical support and valuable discussion. This research was supported by the JSPS KAKENHI (Grant Number: 17H01769, 18J21644, 18H05089).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Grillo, G., Kiang, N.Y. and Moxon, E.C. (1972) Physiological Considerations in Artificial Stimulation of the Inner Ear. *Annals of Otology, Rhinology & Laryngology*, **81**, 714-730. https://doi.org/10.1177/000348947208100513

[2] Fourcin, A.J., Rosen, B.C., Moore, B.C.J., Douek, E.E., Clarke, G.P., Dodson, H. and Bannister, L.H. (1979) External Electrical Stimulation of the Cochlea: Clinical, Psychophysical, Speech-Perceptual and Histological Findings. *British Journal of Audiology*, **13**, 85-107. https://doi.org/10.3109/03005367909078883

[3] Rosen, S. and Ball, V. (1986) Speech Perception with the Vienna Extra-Cochlear Single-Channel Implant: A Comparison of Two Approaches to Speech Coding. *British Journal of Audiology*, **20**, 61-83. https://doi.org/10.3109/03005368609078999

[4] Nie, K., Stickney, G. and Zeng, F.G. (2005) Encoding Frequency Modulation to Improve Cochlear Implant Performance in Noise. *IEEE Transactions on Biomedical Engineering*, **52**, 64-73. https://doi.org/10.1109/TBME.2004.839799

[5] Baskent, D. and Shannon, R.V. (2004) Frequency-Place Compression and Expansion in Cochlear Implant Listeners. *The Journal of the Acoustical Society of America*, **116**, 3130-3140. https://doi.org/10.1121/1.1804627

[6] Peterson, G.E. and Barney, H.L. (1952) Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, **24**, 175-184. https://doi.org/10.1121/1.1906875

[7] Remez, R.E., Rubin, P.E., Pisoni, D.B. and Carrell, T.D. (1981) Speech Perception without Traditional Cues. *Science*, **212**, 947-949. https://doi.org/10.1126/science.7233191

[8] Remez, R.E., Fellowes, J.M., Pisoni, D.B., Goh, W.D. and Rubin, P.E. (1998) Multimodal Perceptual Organization of Speech: Evidence from Tone Analogs of Spoken Utterances. *Speech Communication*, **26**, 65-73. https://doi.org/10.1016/S0167-6393(98)00050-8

[9] Roberts, B., Summers, R.J. and Bailey, P.J. (2010) The Perceptual Organization of Sine-Wave Speech under Competitive Conditions. *The Journal of the Acoustical Society of America*, **128**, 804-817. https://doi.org/10.1121/1.3445786

[10] Feng, Y.M., Xu, L., Zhou, N., Yang, G. and Yin, S.K. (2012) Sine-Wave Speech Recognition in a Tonal Language. *The Journal of the Acoustical Society of America*, **131**, EL133-EL138. https://doi.org/10.1121/1.3670594

[11] Brungart, D.S., Simpson, B.D., Darwin, C.J., Arbogast, T.L. and Kidd Jr., G. (2005) Across-Ear Interference from Parametrically Degraded Synthetic Speech Signals in a Dichotic Cocktail-Party Listening Task. *The Journal of the Acoustical Society of America*, **117**, 292-304. https://doi.org/10.1121/1.1835509

[12] Coath, M., Sheik, S., Chicca, E., Indiveri, G., Denham, S. and Wennekers, T. (2014) A Robust Sound Perception Model Suitable for Neuromorphic Implementation. *Frontiers in Neuroscience*, **7**, 1-10. https://doi.org/10.3389/fnins.2013.00278

[13] Elgendi, M., Bobhate, P., Jain, S., Guo, L., Kumar, S., Rutledge, J., Coe, Y., Zemp, R., Schuurmans, D. and Adatia, I. (2015) The Unique Heart Sound Signature of Children with Pulmonary Artery Hypertension. *Pulmonary Circulation*, **5**, 631-639.

https://doi.org/10.1086/683694

[14] Kondo, T., Amano, S., Sakamoto, S. and Suzuki, Y. (2008) Development of Familiarity-Controlled Word-Lists (FW07). *IEICE Technical Report*, **107**, 43-48.

[15] Tukey, J.W. (1991) The Philosophy of Multiple Comparisons. *Statistical Science*, **6**, 100-116. https://doi.org/10.1214/ss/1177011945

[16] Hillenbrand, J., Getty, L.A., Clark, M.J. and Wheeler, K. (1995) Acoustic Characteristics of American English Vowels. *The Journal of the Acoustical Society of America*, **97**, 3099-3111. https://doi.org/10.1121/1.411872

[17] Lisker, L. (1995) Is It VOT or a First-Formant Transition Detector? *The Journal of the Acoustical Society of America*, **57**, 1547-1551. https://doi.org/10.1121/1.380602

[18] Kluender, K.R. and Lotto, A.J. (1994) Effects of First Formant Onset Frequency on [-Voice] Judgments Result from Auditory Processes Not Specific to Humans. *The Journal of the Acoustical Society of America*, **95**, 1044-1052. https://doi.org/10.1121/1.408466

[19] Benki, J.R. (2001) Place of Articulation and First Formant Transition Pattern Both Affect Perception of Voicing in English. *Journal of Phonetics*, **29**, 1-22. https://doi.org/10.1006/jpho.2000.0128

[20] Sumby, W.H. and Pollack, I. (1954) Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, **26**, 212-215. https://doi.org/10.1121/1.1907309

[21] Goh, W.D., Pisoni, D.B., Kirk, K.I. and Remez, R.E. (2001) Audio-Visual Perception of Sinewave Speech in an Adult Cochlear Implant User: A Case Study. *Ear and Hearing*, **22**, 412-419. https://doi.org/10.1097/00003446-200110000-00005