Perception of Japanese Prosodical Phonemes through use of a Bone-conducted Ultrasonic Hearing-aid

Takayuki KAGOMIYA[1], *Seiji NAKAGAWA*[1]

[1]Health Research Institute,
National Institute of Advanced Industrial Science and Technology (AIST), Japan
{t-kagomiya, s-nakagawa}@aist.go.jp

Abstract

Human listeners can perceive speech signals in a voice modulated ultrasonic carrier from a bone-conduction stimulator, even if the listeners are patients with sensorineural hearing loss. Considering this fact, we have been developing a bone-conducted ultrasonic hearing aid (BCUHA). The purpose of this study is to evaluate the usability of the BCUHA regarding perception of Japanese prosodic phonemes, specifically, perception of singleton/geminate consonants, short/long vowels and pitch accent. For this purpose, a series of phoneme discrimination experiments was conducted. The results of the experiments showed that no significant difference between air-conduction and BCUHA conditions was observed. These results indicate that the BCUHA can transmit short/long vowels, single/geminate consonants and pitch accent information as well as segmental information.

Index Terms: hearing aid, ultrasound, bone-conduction, pitch accent, long vowel, geminate consonant, logistic regression analysis

1. Introduction

For patients with sensorineural profound hearing loss (HL) who are unable to make hearing sensation using a normal hearing aid, we have been developing a bone-conducted ultrasonic hearing aid (BCUHA) [1].

Ultrasound is defined as sound waves that travel at such a high frequency that they cannot be heard by humans. However, if ultrasound is presented through a bone-conducted stimulator (bone-conducted ultrasound, BCU), it can be perceived by human listeners [2] . In addition, if the BCU is amplitude-modulated by speech sound, listeners can perceive the original speech signals [2]. Moreover, this voice-modulated BCU enables sensorineural profound HL patients to perceive speech sounds [1]. The BCUHA which we are developing is based on these phenomena.

The performance of the BCUHA has been evaluated by using mono syllable articulation scores [3] and word intelligibility scores [1]. These studies found that the syllable articulation scores via BCU were over than 60% [3], and the word intelligibility scores for words with high familiarity were more than 85% [1]. The patterns of confusion in speech perception via BCU have many points of similarity to those of air-conduction (AC) [3].

Although the usability of the BCUHA has been evaluated as mentioned above, the evaluations have been restricted to the transmission of segmental information, or textural messages. In other words, little attention has been paid to the transmission of prosody.

For Japanese speakers, perception of suprasegmental phonemes, especially lengths of segments and pitch patterns, is very important because these features are distinctive of the Japanese language. For example, the lengths of vowels /a/ contrast in the words /obasan/ (“aunt”) and /oba:san/ (“grandmother”), the lengths of consonants /t/ contrast in /kita/ (“come”) and /kit:a/ (“cut”). Likewise, /ka’ki/ (“oyster”) and /ka‘ki’/ (“persimmon”) are discriminated by their pitch patterns.

The purpose of this study is to evaluate the performance of perceptions of these suprasegmental phonemes via the BCUHA. It is well known that the discrimination of all minimal pairs as mentioned above is categorical. Thus, the research questions of the present study are as follows: 1) Are the categorized thresholds for identification of the minimal pairs the same as with normal hearing and BCU? 2) Is the categorization sharpness of these discriminations reduced? To answer these questions, several series of psychological experiments were conducted.

2. Method

A categorical peception study generally involves discrimination tasks and identification tasks. In this paper, we address the discrimination tasks. The experiments were designed as identification tasks of each minimal pair which is differentiated by suprasegmental elements as mentioned above. To examine whether differences were observed in the identification between the AC and BCU conditions, the same tasks were conducted under both conditions.

2.1. Design of minimal pairs

All experiments were identification tasks of the stimulus continua. The stimuli were speech sounds that manipulated the target segment lengths or pitch pattern of the original sounds.

2.1.1. Short/long vowel discrimination task

To examine the ability to discriminate between short and long vowels, a minimal pair of nonsense words /etete/ vs. /ete:te/ was selected. The words were set as non-accented words.

The reasons for selecting this minimal pair were as follows. First, as stated above, the words were set as nonsense words. The intelligibility of a word is influenced by the familiarity [5] of the word [4], i.e. the higher the familiarity of a word is, the greater the intelligibility score becomes. Therefore, to avoid this influence, nonsense words were selected.

Second, the segmental sounds of the words are three vowels /e/ separated by two silence plosives /t/ (Figure 1). The reason for selecting /e/ was that /e/ has middle openness. Openness affects the sonority of the vowel: open vowels have high sonor-

ity while close vowels have low sonority. Since the openness of /e/ is middle, the sonority of /e/ is regarded as middle [6]. Therefore, the effect of sonority on intelligibility can be small. Moreover, in Tokyo Japanese, a close vowel placed between two voiceless consonants is regularly devoiced [6, 7]. Since /e/ is a middle vowel, the vowel is rarely devoiced [6, 7]. Furthermore, the target vowel /e/ is placed between two silent plosives /t/, thus manipulation of the vowel is comparatively easy.

Third, the words were designed to avoid influences of Japanese tonal rules. If a vowel has a sharp pitch movement, the vowel is perceived as long [8]. In Tokyo Japanese, a word accent is realized as a sharp pitch fall [9]. This is why the words of the minimal pair were set as non-accented words. In addition, in Tokyo Japanese, if a word is pronounced as an isolated word, a sharp pitch rise appears from the initial mora to the second one unless the initial mora is accented [9]. For example, as shown in Figure 1, the pitch of the second mora /e/ is much higher than that of the first mora. To avoid this effect of initial rising, the target vowel /e/ is placed in the second mora of the three-mora word. Moreover, the mora at the end of the utterance undergoes final lengthening, which makes segmental durations in the word final position longer than others [10, 11]. To avoid this effect, the target mora is placed in the second from the end.

2.1.2. Single/geminate consonant discrimination task

For the singleton/geminate consonant distinction task, a minimal pair of nonsense words /etete/ vs. /etet:e/ was selected. Most of the reasons for selecting this minimal pair are the same as with the long vowel task except for the selection of the target consonant. The consonant /t/ is a silent plosive, so extending and shortening of the consonant is relatively easy (Figure 1). Thus, this is the reason for selecting /t/ as the target consonant.

2.1.3. Pitch accent discrimination task

A minimal pair of real words /a'ka/ ("red") vs. /a'ka'/ ("dirt") was selected.

The reasons for selecting this minimal pair were as follows. First, non-linguistically trained people have great difficulty identifying which mora is accented [12], thus this task was aimed at distinction of a minimal pair of real words.

Second, these words consist of two open vowels /a/ and one voiceless plosive /k/. Since open vowels have good sonority, they are suitable for hearing experiments. In addition, two vowels are separated by a voiceless plosive, thus the manipulation of the pitch of the vowels is easy (Figure 2).

Third, both words have familiarity scores of more than 5.5 [5]. This value is regarded as very high [4]. Thus the effect of familiarity in the detection of each word was assumed to be small.

2.2. Recording of original speech materials

A native Japanese male in his thirties whose native dialect was Tokyo (standard) Japanese participated in the sound recordings.

Each word as mentioned above was spoken by the speaker in an anechoic room. The speaker repeated each word 10 times. The speech samples were recorded at 16 kHz / 16 bit resolution and stored on a personal computer.

2.3. Sound Manipulation

2.3.1. Short/long vowel discrimination task

The stimuli used for the short/long vowel discrimination task were generated in the following manner: An utterance /etete/

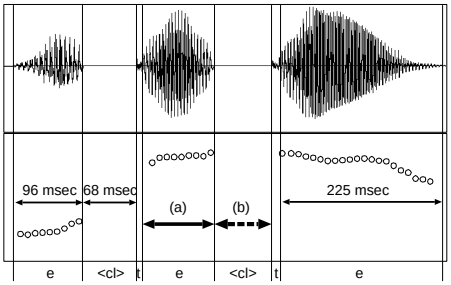

Figure 1: Modification of the target segment of the long vowel discrimination task and the geminate discrimination task. (a) The segment length of /e/ was manipulated for the long vowel task. (b) The closure length of /t/ was manipulated for the geminate task.

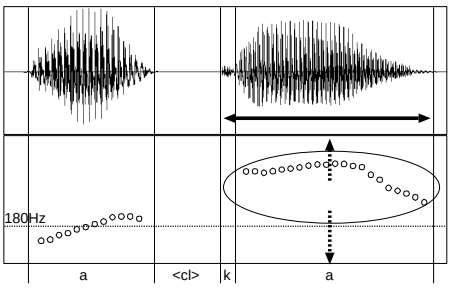

Figure 2: Modification of the stimuli of the pitch accent discrimination task. The original sound was /a'ka'/. The horizontal arrow indicates the time range of /ka/. The average fundamental frequency of /ka/ (circled) was risen or fallen across 180Hz (dotted line).

which had the average length of the target phoneme /e/ in each series of ten repetitions was selected as the original sound. The length of the first phoneme /e/ in the utterance was 96 msec, the second /e/ (target vowel) was 97 msec, and the last /e/ was 225 msec. The target vowel was extended to +200 msec and shortened to -20 msec from the original length by 20 msec step. This range included the minimal length of the vowel in /etete/ (78 msec) and the average (220 msec) and maximum length (234 msec) in /ete:te/ in each series of ten repetitions. The STRAIGHT sound synthesizer [13] was applied for this operation. The STRAIGHT is a tool for manipulating voice quality, timbre, speech length and pitch respectively while keeping natural sound quality. Therefore, the STRAIGHT was suitable for the sound manipulation used in this study. Following to this procedure, a total of 13 stimuli were generated (Figure 1).

2.3.2. Single/geminate consonant discriminate task

An utterance /etete/ was selected applying similar criteria to that of the short/long vowel task. The length of the target closure in the consonant /t/ was 80 msec. This closure length was changed from -20 to +200 msec from the original length in 20 msec steps. This range included the minimal length of the consonant in /etete/ (55 msec) and the average (230 msec) and maximum length (251 msec) in /ete:te/ in each series of ten repetitions. A total of 13 stimuli were generated (Figure 1).

Figure 3: The ceramic vibrator of the BCUHA attached to the mastoid with a hair band-like device

2.3.3. Pitch accent discrimination task

The utterance /a'ka'/, which has average f0 values in both vowels /a/ in each series of ten repetitions was selected as the original sound. The average f0 value in the first /a/ was 155 Hz, and in the second /a/ it was 200 Hz. For this task, the sound modification was raising or lowering of the f0 value of the second mora /ka/ in relation to the first mora /a/. Therefore, the f0 value of /a/ had to be of a fairly intermediate level. Thus, average f0 value of /a/ was risen to 180 Hz. This was about a middle f0 value of /a/ and /ka/ in the original passage. Then, the f0 value of /ka/ was shifted within the range of 130 Hz to 230 Hz in 10 Hz steps. This range was $\pm$50 Hz across 180Hz (Figure 2). This range also included all ten repetitions of /a'ka'/. The STRAIGHT sound synthesizer was also used in this operation. Following this procedure, total of 11 stimuli were generated (Figure 2).

2.4. Experiments

2.4.1. Participants

The participants were ten native speakers of Japanese with no reported speech or hearing defects. Their ages were in the range 22-45 years.

2.4.2. Sound Presentation

In the AC condition, the sound stimuli mentioned above were presented through a headphone (Sennheiser HD650).

In the BCU condition, the presented stimuli were ultrasounds of 30 kHz sinusoid amplitude, modulated by speech signals. The amplitude modulation method applied in this study was the double side band-transmitted carrier (DSB-TC) method, since previous studies have found that this method is capable of spoken language modulation for BCU [1, 3]. With the DSB-TC method, the modulated speech signals $U(t)$ are given by the following expression:

$$U(t) = (S(t) - S_{\min}) \times \sin(2\pi f_c t) \tag{1}$$

where $S(t)$ is the speech signal, $S_{\min}$ is the minimum amplitude value of $S(t)$, and f_c is the carrier frequency (30 kHz).

The BCU stimuli were presented using a custom-made ceramic vibrator (Figure 3). Bone-conducted ultrasound can be perceived when it is applied to various parts of the body, and the mastoids are among the locations where the perception is high. Therefore, we applied the vibrator to the left or right mastoid of the subject using a hair-band-like supporter (Figure 3).

2.5. Procedure

The experiments carried out under both AC and BCU conditions were conducted in a soundproof chamber. Both the presentation of the stimuli and the recording of the responses were executed using a personal computer. Further, the stimuli were provided using a FireWire-based audio interface (Echo Audiofire 12) attached to the personal computer. In both the AC and BCU conditions, the sound levels of the stimuli were adjusted to the most comfortable levels for each participant.

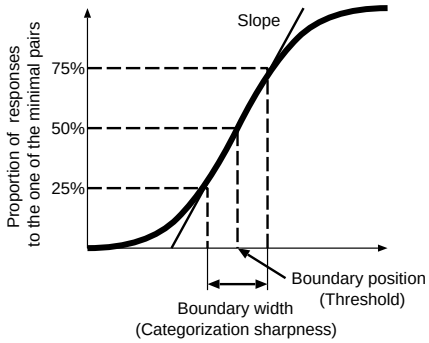

Figure 4: Modeling of categorical identification of minimal pairs by logistic functions.

In the short/long vowel discrimination task, the participants were requested to make a judgment about whether the presented stimulus was /etete/ or /ete:te/. Likewise, the participants judged whether they heard /etete/ or /etette/ in the Single/geminate consonant discrimination task, and made a judgement whether they heard /a'ka'/ or /a'ka/ in the pitch accent discrimination task.

For each participant, two sessions were conducted, with an interval of more than a week between them. The AC condition experiments were conducted in the first session. More than one week after this AC condition session, the BCU condition experiments were conducted.

3. Analysis

Based on the categorical perception theory, the boundary threshold and the sharpness of response were calculated. The proportions of responses can be approximated to logistic function (formula (2)).

$$P(\theta) = \frac{1}{1 + \exp(-k(\theta - \theta_c))} \tag{2}$$

Where θ is the controlled value of the stimuli, and $P(\theta)$ is the proportion of response to the stimuli which have the value θ. θ_c is the θ value at $P(\theta)$ is 50%, k is the slope of the logistic curve in θ_c, i.e. the maximum value of the slope [14].

All parameters were calculated applying the least square method. Although k is the parameter of categorization sharpness, it is difficult to understand immediately how the slope value indicates sharp categorization, thus each θ in which $P(\theta)$ became 25% and 75% was calculated, the range between these two θ was defined as the parameter of categorization sharpness (Figure 4).

The proportion of responses to each alternative set of minimal pairs was calculated for each participant. By letting manipulated segment lengths or pitch heights be θ, and letting the proportion of responses to the stimulus θ be $P(\theta)$, the parameters of the logistic function above were calculated.

As a sequel of this procedure, the thresholds of categorization and sharpness of categorization to each minimal pair by each participant were calculated.

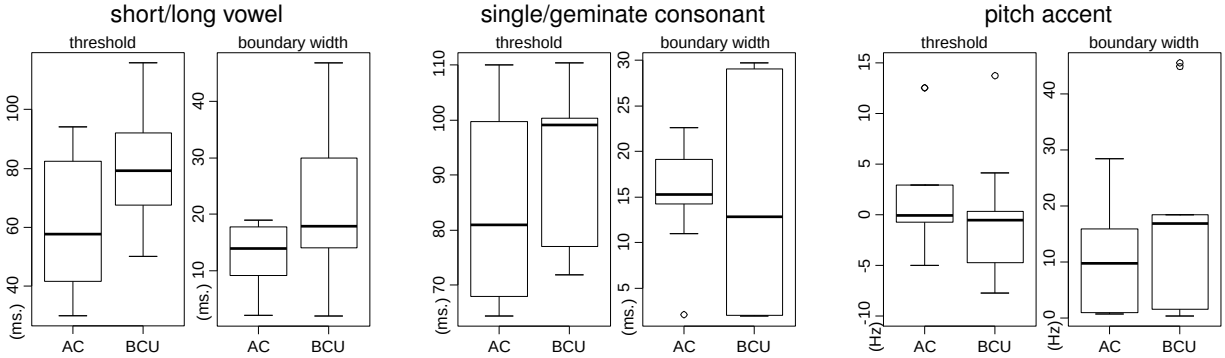

Figure 5: Distribution of participants for threshold and boundary width of categorization for each minimal pair.

4. Results and Discussion

Figure 5 shows the distribution of participants for the threshold and categorization range. The ordinates in the graphs of the threshold indicate the distance from the original sounds to the categorical boundaries.

In each task, a series of paired t-tests revealed that both the thresholds and ranges of the AC condition and the BCU condition did not differ significantly at the 0.05 level. This result suggests that each suprasegmental phoneme can be discriminated even through BCU.

At the same time, although there were no significant differences, there were tendencies that the thresholds of single/geminate consonant discrimination task and short/long vowel discrimination task shifted toward the later and longer side (short/long vowel discrimination task: $p = 0.074$, single/geminate consonant discrimination task: $p = 0.127$). This tendency suggests that BCU discrimination of long segments from short ones requires a greater difference between them than is the case with AC. What is more, there were tendencies that the individual difference of the ranges became larger in the single/geminate consonant discrimination task and short/long vowel discrimination task. To investigate whether the difference of threshold between AC and BCU had any relation to the difference of range between AC and BCU, the correlation coefficient between the two differences was calculated. The result revealed that there was a high correlation coefficient ($r = 0.745$, $p < 0.05$) in the single/geminate discrimination task. This correlation suggests that the participants who is inferior in single/geminate discrimination task in BCU condition require more differences between geminate and single consonant to recognize the consonant is geminate.

5. Conclusion

This research examined the performance of discrimination in suprasegmental phonemes. The following were the main findings; 1) The threshold of discrimination to categorize suprasegmental phonemes in Japanese did not change with BCU. 2) Also the sharpness of discrimination was not reduced with BCU. Overall, these findings suggest that the BCUHA will assist users in hearing languages even if suprasegmental features are distinctive in the languages.

However, there were individual differences among the participants in terms of their discrimination abilities with BCU. Further study should be carried out to discover why these individual differences occurred.

6. Acknowledgments

This research was supported by the Funding Program for Next-Generation World-Leading Researchers provided by the Cabinet Office, Government of Japan. 2009FY Grants-in-Aid for Scientific Research of Japan Society for the Promotion of Science (21700592) , and the 2009FY Research Grant of the Sound Technology Promotion Foundation of Japan.

7. References

[1] S. Nakagawa *et al.*, "Development of a bone-conducted ultrasonic hearing aid for the profoundly sensorineural deaf," *Trans. of the Japanese Soc. for Medical and Biological Engineering : BME*, vol. 44, no. 1, pp. 184–189, 2006.

[2] M. L. Lenhardt, *et al.*, "Human ultrasonic speech perception," *Science*, vol. 253, pp. 82–85, 1991.

[3] Y. Okamoto, *et al.*, "Inteligibility of bone-conducted ultrasonic speech," *Hearing Research*, vol. 208, pp. 107–113, 2005.

[4] S. Sakamoto, *et al.*, "Complementary relationship between familiarity and SNR in word intelligibility test," *Acoust. Sci. and Tech.*, vol. 25, no. 4, pp. 290–292, 2004.

[5] S. Amano and T. Kondo, *Lexical Properties of Japanese*. Tokyo: Sanseido, 1999, vol. 1.

[6] H. Kubozono, *Nihongo no Onsei [Phonetics of Japanese]*. Tokyo: Iwanami Shoten, 1999.

[7] J. D. McCawley, *The Phonological Component of a Grammar of Japanese*. The Hague: Mouton, 1968.

[8] I. Takiguchi, *et al.*, "Effects of a dynamic f0 on the perceived vowel duration in japanese," in *Proc. of Speech Prosody 2010*, 2010, pp. 944:1–4.

[9] J. B. Pierrehumbert and M. E. Beckman, *Japanese Tone Structure*. The MIT Press, 1988.

[10] D. H. Klatt, "Vowel lengthening is syntactically determined in a connected discourse," *J. of Phonetics*, vol. 3, pp. 129–140, 1975.

[11] K. Takeda, Y. Sagisaka, and H. Kuwabara, "On sentence-level factors governing segmental duration in Japanese," *J. Acoust. Soc. Am.*, vol. 86, no. 5, pp. 2081–2087, 1989.

[12] O. Mizutani, "Akusento to intoneshon no shutoku-hou [learning method for accent and intonation]," in *Nihongo no onsei, on'in (Ge) [Japanese Phonetics and Phonology II]*, ser. Kouza Nihongo to Nihongo kyouiku [Japanese and Japanese Education cource], M. Sugito, Ed. Tokyo: Meiji Shoin, 1990.

[13] H. Kawahara, "STRAIGHT, exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoust. Sci. and Tech.*, vol. 27, no. 6, pp. 349–353, 2006.

[14] F. B. Baker, *The Basics of Item Response Theory*, 2nd ed. University of Maryland, College Park, MD: ERIC Clearinghouse on Assessment and Evaluation, 2001.