The Effects of Fundamental Frequency and Formant Space on Speaker Discrimination through Bone-conducted Ultrasonic Hearing

Takayuki KAGOMIYA¹, Seiji NAKAGAWA¹

¹Institute for Human Science and Biomedical Engineering, National Institute of Advanced Industrial Science and Technology (AIST), Japan
{t-kagomiya, s-nakagawa}@aist.go.jp

Abstract

Human listeners can perceive speech signals from a voice-modulated ultrasonic carrier which is presented through a bone-conduction stimulator, even if they are sensorineural hearing loss patients. As an application of this phenomenon, we have been developing a bone-conducted ultrasonic hearing aid (BCUHA). This research examined whether formant space and F₀ can be cues of speaker discrimination in BCU hearing as well as via air-conduction (AC) hearing. A series of speaker discrimination experiments revealed that both formant space and F₀ can act as cues for speaker discrimination even via BCUHA. However, sensitivity to formant space in BCU hearing is less than in AC hearing.

Index Terms: ultrasound, bone-conduction, hearing aid, speaker discrimination, multi-dimensional scoring

1. Introduction

To provide sound sensations for sensorineural hearing loss patients, we have been developing a bone-conducted ultrasonic hearing aid (BCUHA) [1]. This hearing aid is based on the phenomenon that ultrasonic vibration generates a sensation of sound by bone-conduction. In addition, this bone-conducted ultrasound (BCU) can transmit speech sounds. Amplitude-modulated BCU by speech sounds are perceived as original speech sounds [2]. Moreover, the amplitude-modulated bone-conducted ultrasonic speech sounds can be perceived by sensorineural hearing loss patients [1].

In terms of the usability of BCUHA, Japanese syllable articulation and word intelligibility have been evaluated [3]. The results showed that 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 BCUHA has been evaluated as mentioned above, the evaluations have been restricted to transmission of linguistic information. Little attention has been paid to transmission of paralinguistic or nonlinguistic information.

The purpose of this study is to evaluate the usability of BCUHA in respect of paralinguistic information transmission, especially transmission of information concerning speaker discrimination or identification. To maintain good communication, speaker discrimination or identification plays a crucial role. Therefore, it is important to evaluate the performance of speaker discrimination in BCU hearing.

Listeners use many acoustical or linguistic cues for speaker discrimination, for example, pitch, speaking rate, formant space, vocal tension, vocabulary choice, and so on. In this research, we focused on pitch and formant space.

Pitch is a very important factor for speaker discrimination or identification. It is well known that the voices of adult males have a low pitch, while the voices of females and children have a high pitch. This tendency is derived from the difference in the larynx size and vocal fold length [4]. Similarly, formant space size reflects vocal tract length. Adult males have long vocal tracts, children have short ones, and those of adult females are medium length. Corresponding to these vocal tract length differences, formant space sizes of adult males are narrow, children’s are wide, and females’ are middle-sized [5]. Furthermore, this correspondence between physical and acoustical features is used to provide perceptual cues to judge speakers’ genders or ages [6]. Low pitch and narrow formant space voices are perceived to indicate voices of adult males, while high pitch and wide formant space voices are judged to be voices of children [6]. Consequently, pitch and formant space are significant cues for speaker identification or discrimination.

However, auditory sensitivities for formant spaces and pitches via BCU hearing have not been evaluated. As for BCU hearing, formant domain resolution is not so high as AC hearing, especially under 250 Hz and over 8 kHz [7]. This non-linearity of formant domain resolution may cause variation in sensitivity to change of formant spaces and pitches.

In this study, to examine whether pitch and formant space can act as cues to speaker discrimination even through BCU hearing, we conducted a series of listening experiments using speech sounds with controlled pitch and formant space. In addition, it is expected that the results of this experiments will provide clues to understanding the characteristics of the formant domain sensitivity of BCU hearing.

2. Method

Participants were presented with a series of isolated words in pairs. The words were manipulated in F₀ (pitch) and formant space respectively. Participants were requested to make a judgment about whether the speakers of the paired words were the same or different. To examine whether there were differences between AC and BCU hearing in speaker discrimination, the same tasks were conducted in AC and BCU conditions.

2.1. Stimuli

2.1.1. Original material

It has been reported that identification of familiar speakers can be easier than is the case for unfamiliar speakers [8]. To avoid this effect, it is preferable that the speaker of the stimuli is unfamiliar to the participants. Thus the original material was selected from corpus. To select an average or typical speaker, we used “The Corpus of Spontaneous Japanese” (CSJ) [9] in which a large number of speakers’ voices were stored.
Also, the original material used in this research was not a male voice but a female’s. The reason for using a female voice was that male voices have a low pitch, therefore scaling down the pitch might make the voice too unnatural.

An additional requirement for selecting the phrase pronounced by a typical speaker is that the same phrase must be pronounced by a large number of persons. Thus, the material was selected from short passage reading tasks stored in CSJ. CSJ contains two short passage reading tasks. For this research, the first phrase “Uchu:” (the universe) in the “Uchu:” task was used.

The average female speaker’s utterance was selected in the following manner. A total of 123 females participated in the “Uchu:” task. For each speaker’s utterance, the mean values of F0 and speaking rate were computed. The AMDF method of the Snack Sound Toolkit (http://www.speech.kth.se/snack/) was used for F0 estimation (10 ms step). Then mean values for total speaker’s F0 and speaking rate were calculated. Finally, an utterance nearest to the mean values of all the speakers was selected. The F0 value of the selected utterance was 127.871 Hz, and the speaking rate value was 6.315 moras per second.

### 2.1.2. Sound manipulation for making stimuli

Stimuli were created from the original sound by a series of manipulations of its F0 and formant space. As for F0, three levels of F0 values were adapted, i.e. 20% lower (102.297 Hz), 20% higher (153.445 Hz), and the original value. The lower and higher values are almost the mean value ±1SD in female voices of the “Uchu:” dataset, thus neither value was an unusual level. Moreover, the lower value was lower than the mean ±1SD in male voices in the equivalent part of CSJ (104.391 Hz), thus the value was not an unusual level even in a male voice.

Also the formant space was stretched by 20%, shrunk by 20%, and kept at the original size. These values were determined to represent the ratio of vocal tract length of males to that of females. As mentioned above, a person’s formant space size reflects the vocal tract length. According to source-filter theory, each formant (Fn) can be predicted using the following equation:

\[
F_{n} = \frac{(2n-1)c}{4f}\tag{1}
\]

where \(c\) is the speed of sound (34400 cm/sec.), and \(f\) is vocal tract length. Accordingly, vocal tract length can be estimated from formant frequencies using inverse operations of the equation. Thus the vocal tract length (VTL) was estimated from measured formant frequencies using the following procedure.

1. F1, F2, F3 were calculated using the LPC method of the Snack Sound Toolkit (12th order, 10 ms window).
2. Vocal tract lengths \(l\) were estimated from each \(F_{n}\) using the following formula (reverse operation of equation (1.1)):

\[
l = \frac{(2k+1)c}{4F_{k+1}}\tag{2}
\]

where \(k = (0,1,2), F_{k} + 1\) is the formant frequency of interest.
3. The mean of \(l\) calculated from the three formants was regarded as the estimated vocal tract length (VTL).

### Figure 1: DSB-TC modulation of the sound

The VTL of the original sound was 14.265 cm. This value is slightly greater than actual values measured using MRI [10]. A reasonable explanation is that the vowel “u” has a lower F1 and F2, i.e. if a lower formant frequency is applied to equation (2), VTL has a greater value than a neutral vocal tube.

The VTL 20% longer than the original size is 17.119 cm. This value was within the mean ±1SD in the value of male VTL in the corpus. On the other hand, a 20% shorter size is 11.412 cm. This value is almost the average size for VTL of low-teenagers [10].

Counting the results of each of the three levels of manipulation in F0 (Low, Original, High — L, O, H) and formant space (Narrow, Original, Wide — N, O, W), a total of nine stimuli were generated.

All sound manipulation processes described above were carried out by using the STRAIGHT speech synthesizer [11], which enables to separate source information from the spectral-envelope information and manipulate respectively.

### 2.2. Participants

Nine native speakers of Japanese with no reported speech or hearing defects participated in the experiments.

### 2.3. Sound Presentation

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

On the other hand, 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 found that this method is capable of spoken language modulation for BCU [3, 7]. With the DSB-TC method, the modulated speech signals \(U(t)\) are given by the following expression:

\[
U(t) = (S(t) - S_{\text{min}}) \times \sin(2\pi f_{c}t)\tag{3}
\]
Table 1: The number of responses in which the listeners judged that the speakers in the pairs of interest were “different”

<table>
<thead>
<tr>
<th>Pair</th>
<th>AC</th>
<th>BCU</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-N</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>L-O</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>L-W</td>
<td>81</td>
<td>64</td>
</tr>
<tr>
<td>O-N</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>O-O</td>
<td>72</td>
<td>65</td>
</tr>
<tr>
<td>O-W</td>
<td>81</td>
<td>64</td>
</tr>
<tr>
<td>H-N</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>H-O</td>
<td>56</td>
<td>64</td>
</tr>
<tr>
<td>H-W</td>
<td>81</td>
<td>64</td>
</tr>
</tbody>
</table>

where \(S(t)\) is the speech signal, \(S_{\text{min}}\) is the minimum amplitude value of \(S(t)\), and \(f_i\) is the carrier frequency (30 kHz) (Figure 1).

The sampling frequency of the original signal and the bandwidth of the DSB-TC modulated signals were 16 kHz. Since the carrier frequency was 30 kHz, all frequency components included in the DSB-TC modulated signals were over 20 kHz.

The stimuli of the BCU condition were presented using a custom-made ceramic vibrator [1] (Figure 2). Bone-conducted ultrasound can be perceived when it is delivered to various parts of our body, and the mastoid is one of the best locations for perception. Therefore, we applied the vibrator to the left or right mastoid of the subject using a hair band-like supporter.

### 2.4. Apparatus

Both the presentation of the stimuli and the recordings of the responses were executed using a personal computer. Also the stimuli were played using a FireWire-based audio interface (Echo AudioFire12) attached to the personal computer.

The sound levels of the stimuli were adjusted to the most comfortable levels for each participant.

In both the AC and BCU conditions, the experiments were conducted in a soundproof chamber.

### 2.5. Procedures

Each stimulus was presented as a pair, and participants were requested to make a judgment about whether speakers were the “same” or “different”. Each pair was presented 10 times. The overall order of presentation of the pairs was also randomized. Each listener participated in the BCU experiments first, and a few days later took part in the AC test.

### 3. Analysis

Table 1 presents the number of responses in which the listeners judged that the speakers in the pairs of interest were “different”. The upper table shows the AC condition, the lower is the BCU condition. The responses of all participants were pooled. Since nine listeners responded 10 times to each pair, the maximum possible number was 90.

According to Table 1, diagonal elements have quite small values in both the AC and BCU conditions (max. 6). This result indicates that the sound pairs identical in both the F0 and formant spaces were perceived as being the same speaker’s voice, not only in the AC condition, but also in the BCU condition.

To obtain a broader view of the response tendency to the stimuli, a series of multi-dimensional scaling (MDS) was conducted. As the response data shown in Table 1 were regarded as the psychological distance between each sound, Kruscal’s non-metric multi-dimensional scaling (NMDS) was adopted for both the results of the AC and BCU conditions. To apply NMDS, each distance between the same stimulus (diagonal elements in Table 1) was substituted for zero. This operation was required for NMDS calculation, but was not unfair because the distances between the same stimuli were quite small, as mentioned before.

Stress values were checked to decide appropriate numbers of dimensions. Each stress value revealed that three dimensions were sufficient for both the results of AC (0.005) and BCU (0.054).

### 4. Results

Figure 3 shows the distribution of each stimulus by the results of NMDS. The shapes of the points indicate the F0 value, and the colors of the points represent the formant space size. In the case of AC, three tight clusters were created in the 1-2 dimension. Each cluster was grouped according to formant spaces. While in the 2-3 dimension, the stimuli were located by the order of the formant sizes on dimension 2, and placed according to the F0 value on dimension 3. Although the distribution in dimension 3 indicates that differences of the F0 value could also be perceived, nevertheless dimension 3 was not important because the range of dimension 3 was small and the stress value of dimension 3 was also quite small, as mentioned before. These results indicate that formant space is a primal factor for speaker identification.

On the other hand, such sharp-formed clusters did not occur in BCU. However, there were clear tendencies in the 1-2 dimension in BCU. In dimension 1, every stimulus was distributed according to the F0 value, while in dimension 2, each sound was ordered by the size of the formant spaces. Moreover, in the 2-3 dimension, the stimuli were loosely grouped by the formant spaces, and in the 1-3 dimension, they were loosely grouped by the F0 value. These results revealed that both the F0 value and formant space can be perceived as speaker identification factors even in BCU hearing.

### 5. Discussion

The results of speaker discrimination experiments showed that changes of formant space size and F0 value are perceived not only in AC hearing but also in BCU hearing. However, some differences in response tendencies between AC and BCU hearing were observed.

In the AC condition, formant space size was a dominant factor for speaker discrimination. Although the influence of F0 manipulation was quite small, differences of F0 could also be perceived. A possible reason for this is the different sources from which both acoustical parameters derived. The formant space size was mainly derived from the speaker’s vocal tract length, while differences of the F0 value were derived from glottal pulse rates. The results of the AC experiments suggest that speaker discrimination depends on formant space size because differences in formant space sizes are reflections of speakers’ physical differences in vocal tract lengths. On the other hand, differences in the F0 value are regarded as speech variations within each speaker. Also, these results are consistent with acoustical analyses of F0 and formants concerning speaker
identities [12].

In contrast to the result of the AC experiment, formant space size was not a dominant factor in BCU hearing. Differences of formant space size and F0 value contributed equally to speaker discrimination. In other words, the listeners had similar sensitivities to differences of F0 value and differences of formant spaces. It is more likely that this tendency can be explained by a reduction in the sensitivity to formant space difference in BCU hearing, rather than an increase in the sensitivity to of the F0. A study of frequency resolution of BCU hearing revealed that difference limens for frequency (DLF) through BCU is increased under 250 Hz and over 8 kHz [7]. It is reasonable to assume that this growth of DLF causes lower sensitivity to differences in formant spaces.

However, although reductions in sensitivity were found, there are still sensations of formant space in BCU hearing. As indicated in Figure 3, differences of formant spaces were reflected in psychological spaces. Therefore, appropriate improvements in frequency resolution of BCU hearing can make formant spaces a more prominent factor for speaker identification.

6. Summary and Conclusions

The effects of formant space and F0 on speaker discrimination through BCU hearing were examined. For this purpose, a series of speaker discrimination experiments were conducted. The stimuli were synthesized voices in which formant space and F0 were modified. The results show that both formant space and F0 can be clues for speaker discrimination even in BCU hearing; however, sensitivity to formant space is less than is the case with AC hearing. To solve this problem, further investigations are required.

7. Acknowledgments

This research was supported by the Industrial Technology Research Grant Program of the New Energy and Industrial Technology Development Organization Japan, the Strategic Information and Communications R&D Promotion Programme of the Ministry of Internal Affairs and Communications Japan, the Grants-in-Aid for Scientific Research of Japan Society for the Promotion of Science, and the research grant of the Sound Technology Promotion Foundation of Japan.

8. References


