Study on Effects of Speech Production during Delayed Auditory Feedback for Air-Conducted and Bone-Conducted Speech

Teruki Toya1, Daisuke Ishikawa1, Ryota Miyauchi1, Kazushi Nishimoto2 and Masashi Unoki1

1 School of Information Science, Japan Advanced Institute of Science and Technology
1–1 Asahidai, Nomi, Ishikawa 923–1292, Japan
E-mail: {yattin_yatson, dishikawa, ryota, unoki}@jaist.ac.jp

2 School of Knowledge Science, Japan Advanced Institute of Science and Technology
1–1 Asahidai, Nomi, Ishikawa 923–1292, Japan
E-mail: knishi@jaist.ac.jp

Abstract

To investigate the relationships between speech perception and production, speaking styles and performances of speakers have been investigated during speech production under delayed auditory feedback (DAF). However, previous studies have focused on only the delay in feedback for air-conducted speech although speakers perceive their own voice for both air-conducted and bone-conducted speech. In this paper, the phenomena between speech production and perception were investigated under DAF presented as both air-conducted and bone-conducted speech. It was then confirmed whether the speaking styles and performances were similar or different for these two types of presentation. The ratio of speech duration under delay conditions to that under a non-delay condition ($R_d$) and the number of dysfluent episodes ($N_d$) were measured to quantify the effect of delayed speech on the speaking styles and performances. As a result, long duration and several dysfluencies were observed under DAF for both air-conducted and bone-conducted speech. Moreover, the changes under DAF for bone-conducted presentation tended to be larger than those for air-conducted presentation. These results suggest that the effect of delay in feedback due to bone-conducted speech might be stronger than that due to air-conducted speech.

1. Introduction

Speakers perceive their own voices to control their production systems while communicating through speech [1]. This mechanism is referred to as auditory feedback. There is a technique of feeding speakers’ delayed voice back to themselves, called “delayed auditory feedback (DAF).” DAF can be used to investigate the relationships between speech production and perception because their speaking styles and performances are affected by the delay when perceiving their delayed voice.

Previous studies have found that longer duration and dysfluencies, such as prolongation and repetition, are observed in the 50–200 ms delay range, while they are not observed much above 300 ms delay under DAF [2], [3]. Surprisingly, stutterers’ speech tends to become more fluent under DAF [4], [5]. However, these studies have focused on only the delay in feedback due to air-conducted speech and ignored effects of that due to bone-conducted speech. Practically, speakers’ own voices are fed back as both air-conducted and bone-conducted speech. The phenomena stated above might occur from delays for not only air-conducted but also bone-conducted speech.

This study aims to investigate the phenomena between speech production and perception under DAF presented as both air-conducted and bone-conducted speech and carefully consider whether they are similar or different.

2. Method

2.1 Participants

Four male speakers aged 22 to 26 participated in the experiments. All speakers had normal hearing and none had speaking disorders. They were native Japanese speakers.

2.2 Apparatus

Figure 1 shows a schematic diagram of DAF for this study. Experiments were conducted in a soundproof room. Speakers’ voices were recorded through a microphone (Sennheiser MKH 416 P48) and routed through an audio interface (Roland QUAD-CAPTURE) to a PC (LG Sharkoon, with OS Ubuntu Studio 14.04 Linux). The sampling frequency was 44,100 Hz and the number of quantizing bits was 16. The speech signal was temporally delayed by the PC depending on the experimental conditions. The signal was routed through the interface, amplified by an audio amplifier (audio-technica AT-HA5000) and presented to the speakers through a headphone (Sennheiser HD280) or a bone-conduction transducer (TEAC HP-F200). The total latency in the system was 13.2 ms.

2.3 Procedure

Speakers uttered three Japanese words consisting of four morae (/sjabusjabu/, /takabisja/, /yuuwaku/). They uttered each word one by one under a non-delay condition. Under this condition, the utterances were repeated five times per word. They then uttered each word under six delay conditions...
(25, 50, 100, 150, 200, and 400 ms). The delay conditions were randomly changed while the utterances were repeated 30 times per word. In the end, each delay condition contained five utterances. These processes were then repeated twice (one for the air-conducted and one for the bone-conducted presentation). For both air-conducted and bone-conducted presentation, no other sound or noise was presented to the speakers. For bone-conducted presentation, earplugs were inserted into their ears to prevent them from hearing external sound by air-conduction.

At the beginning, the feedback intensity was adjusted to a level comfortable for each speaker and fixed. During the experiment, speakers were required to maintain as much of their vocal intensity as possible.

2.4 Analysis

To investigate the effect of delayed speech on the speaking style, the ratio of speech duration under a delay condition to that under the non-delay condition \(R_d\) and the total number of dysfluent episodes \(N_d\) were determined as a function of delay.

The \(R_d\) is defined in Eq. (1), where \(T_i\) [ms] is the \(i\)-th speech duration under the non-delay condition, \(T_{d,i}\) [ms] is the \(i\)-th speech duration under each delay condition \((d = 25, 50, \cdots, 400)\), and \(N_w\) is the number of one-word utterances per delay condition for each speaker \((N_w = 5, i = 1, 2, \cdots, N_w)\).

\[
R_d = \frac{1}{N_w} \sum_{i=1}^{N_w} \frac{T_{d,i}}{T_i}
\]

\[
T = \frac{1}{N_w} \sum_{i=1}^{N_w} T_i
\]

Figure 2 shows examples of \(T_i\) and \(T_{d,i}\) in speech waveforms for both air-conducted and bone-conducted presenta-
tion. The $T_i$ and $T_{d,i}$ were calculated from the starting and end points in the waveforms, represented as red solid lines.

A part-word prolongation or part-word repetition was defined as the dysfluent episode from recorded voices of speakers, in the same way as a previous study [3]. In this study, it was defined by one experimenter although in the previous study it had been defined by several experimenters. The total number of dysfluent episodes ($N_d$) was then defined using Eq. (3), where $N_c$ is the total number of utterances per delay condition for all four speakers and all three words ($N_c = 60, j = 1, 2, \cdots, N_c$).

$$N_d = \sum_{j=1}^{N_c} s_j$$

where

$$s_j = \begin{cases} 1, & \text{Dysfluent episodes were observed} \\ 0, & \text{Dysfluent episodes were not observed} \end{cases}$$

Figure 2 also shows examples of $s_j$ in speech waveforms. In the waveforms under the non-delay condition, no dysfluent episodes were observed ($s_j = 0$). In the waveforms under the delay conditions, some dysfluent episodes were observed and are indicated as a part-word prolongation (Fig. 2(a)) and a part-word repetition (Fig. 2(b)) ($s_j = 1$), represented as magenta circles.

3. Results

3.1 Speech duration

Figure 3 shows the $R_d$ of the word /sjabusjabu/ spoken by speaker-A, /takabisja/ by speaker-B, and /yuuwaku/ by speaker-C as a function of delay.

The increase in $R_d$ was observed for both air-conducted and bone-conducted presentation, depending on the delay conditions, as shown in Figs. 3(a) and 3(b). For air-conducted presentation, $R_d$ tended to increase gradually and become peak in the 100–200 ms delays. At the same time, no large increase/decrease in $R_d$ was observed as the delay exceeded 200 ms. For bone-conducted presentation, however, $R_d$ tended to increase rapidly at 25 ms delay. Then, the delay at which $R_d$ became peak tended to be different among the speakers. Overall, it was found that $R_d$ for bone-conducted presentation tended to be a little larger than that for air-conducted presentation.

Moreover, the contours of each $R_d$ as a function of delay tended to be different among the types of uttered words. In particular, $R_d$ and its standard deviation of the word /sjabusjabu/ (Fig. 3(a)) tended to be larger than any other words.

In speaker-C (Fig. 3(c)), $R_d$ did not largely change among the delay conditions, for both air-conducted and bone-conducted presentation.

3.2 Speech dysfluencies

Figure 4 shows the $N_d$ as a function of delay. The increase in $N_d$ was observed for both air-conducted and bone-conducted presentation, depending on the delay condition. For air-conducted presentation, $N_d$ tended to be a little larger
in the 50–200 ms delays than under the other delay conditions. For bone-conducted presentation, however, \( N_d \) tended to increase rapidly, becoming peak at the 200 ms delay and decreased rapidly as the delay exceeded 200 ms. Overall, it was found that \( N_d \) for bone-conducted presentation tended to be larger than that for air-conducted presentation.

Dysfluent episodes were observed more frequently in the word /sjabusjabu/ (32 episodes) than any other words (8 episodes). Moreover, the number of dysfluent episodes differed among the speakers (A: 13, B: 0, C: 7, and D: 20).

4. Discussion

The experimental results are summarized as follows:

1. Several changes in speaking style and performance, such as longer duration and dysfluencies, were observed during speech under DAF for both air-conducted and bone-conducted presentation.

2. Overall, larger changes were observed under DAF for bone-conducted presentation than for air-conducted presentation.

These results indicate that not only air-conducted but also bone-conducted feedback affect one’s speaking style and performance. In addition, considering the second result, the effect of delay in feedback for bone-conducted speech might be stronger than that for air-conducted speech. It may be assumed that perceiving one’s own voice via bone-conduction might have great importance in controlling one’s own production system.

The speaking styles and performances for a speaker did not largely change regardless of the delay conditions or presentation types (air-conducted/bone-conducted speech). In fact, the effects of delayed voice on one’s speech seem to be different among speakers. Moreover, their speaking styles and performances might change gradually while they become accustomed to the experimental conditions.

There is a remaining issue to be considered: It seems to be difficult to separate air-conducted/bone-conducted speech from a speaker’s voice. It has been assumed that there are many paths for a speaker’s voice to be transmitted among his/her skin, bones, and auditory system [6].

For further work, to reveal the relationships between speech perception and production, it will be investigated how auditory feedback due to air-conducted and bone-conducted speech separately works for speech communication by considering the transmission properties of both air-conducted and bone-conducted speech.

5. Conclusion

In this paper, the phenomena between speech production and perception were investigated under DAF presented as both air-conducted and bone-conducted speech. It was then confirmed whether the speaking styles and performances were similar or different for these two types of presentation. The ratio of speech duration under delay conditions to that under a non-delay condition \( (R_d) \) and the number of dysfluent episodes \( (N_d) \) were measured to quantify the effect of delayed speech on the speaking style. As a result, long duration and several dysfluencies were observed under DAF for both air-conducted and bone-conducted presentation. Moreover, the changes under DAF for bone-conducted presentation tended to be larger than those for air-conducted presentation. These results suggest that the effect of delay in feedback due to bone-conducted speech might be stronger than that due to air-conducted speech.

Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research (A) (No. 25240026) and Young Scientists (A) (No. 24683026).

References