Applied Acoustics 140 (2018) 263–274

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Effects of inter-word pauses on speech intelligibility under long-path echo conditions

Shuichi Sakamoto[a,b,*], Zhenglie Cui[a,b], Tomori Miyashita[a,b], Masayuki Morimoto[a], Yôiti Suzuki[a,b], Hayato Sato[c]

[a] Research Institute of Electrical Communication, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai, Miyagi 980-8577, Japan
[b] Graduate School of Information Science, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai, Miyagi 980-8577, Japan
[c] Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan

ARTICLE INFO

Keywords:
Word intelligibility
Long-path echo
Pause
Open-air public-address systems

ABSTRACT

Long-path echo is a salient factor that causes the degradation of the intelligibility of speech transmitted through a wide area outdoor environment or a very large indoor space using public-address systems. To robustly transmit speech information under such conditions, it is important to overcome this effect by controlling the characteristics of speech sounds. In this study, we consider the effects of inserting pauses between the words of a sentence. We performed word intelligibility tests using a series of four continuous words, called a quadruplet. Various pause lengths and long-path echo patterns were applied to the quadruplet. The results of the experiments demonstrate that word intelligibility under a long-path echo is significantly improved by the insertion of pauses between the words. Intelligibility can approach the same levels observed in the absence of echoes for a pause length of approximately 200 ms, which is almost the same as the length of 1-mora for the words used in the experiments. Moreover, this 200 ms pause is known to be sufficient to improve speech recognition in older adults. These results suggest that inter-word pauses of a length of approximately 1-mora can generally enhance the robustness of speech communication systems when used under a severe environment.

1. Introduction

When public-address systems convey speech information to a wide outdoor area or to a very large indoor space, numerous factors cause degradation of the intelligibility of the sound produced by the systems. As speech is transmitted over long distances, high-frequency components, which are important constituents to distinguish consonants, are attenuated by air [1]. In reverberant conditions such as the platform of a railway station, echo sounds as well as reverberation sounds cause the degradation of the intelligibility of announcements [2,3]. Background noise is also an important factor to be considered. However, in outdoor environments or in large indoor spaces, the effect of long-path echoes become crucially important.

Long-path echoes are generated by reflections from mountains, buildings, and other large-size surroundings. This phenomenon is often observed in Japan, since most Japanese local governments have installed acoustic mass notification systems using emergency outdoor public-address systems. These systems are mostly connected to governmental radio communication networks. Outdoor public-address systems can effectively and simultaneously transmit various information over a wide service area. Since the audio output from several outdoor public-address systems are often received at a listening point, long-path echoes are generated not only by the reflections mentioned above, but also by the same sound from neighboring systems. This means that those "direct sounds" from neighboring systems that are not the nearest systems to the listening point can also be regarded as long-path "echoes." The delay times correspond to the difference in the distance between the nearest system and the other systems. Therefore, the delays are often more than a couple of hundred milliseconds, which is much longer than those observed with public-address systems in a large room. Thus, it is crucially important to investigate various methods of conveying spoken information to an entire service area with sufficient intelligibility under long-path echo conditions. This investigation should be based on a good understanding of the effects of long-path echoes on speech intelligibility.

Nevertheless, only a few studies are available on the effect of long-path echoes on speech intelligibility in outdoor environments, with the notable exception of several works by Toida [4–6]. These studies highlighted that speech intelligibility is degraded by long-path echoes, and that this degradation can be determined using masking [6].

* Corresponding author at: Research Institute of Electrical Communication, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai, Miyagi 980-8577, Japan.
E-mail address: saka@ais.riec.tohoku.ac.jp (S. Sakamoto).

https://doi.org/10.1016/j.apacoust.2018.01.020
Received 10 November 2017; Received in revised form 15 January 2018; Accepted 19 January 2018

Recently, our research group determined that speech intelligibility becomes high regardless of the existence of long-path echoes under the condition in which the speech transmission index (STI) is higher than 0.6. We also discovered that long-path echoes degrade speech intelligibility under low-STI listening conditions [7]. The occurrence of these low-STI areas are inevitable in outdoor public-address systems, because the delay times and consequently the STI values, depend on the relative distances between the listening point and the source of the echoes.

To transmit speech information under such severe listening conditions, several methods have been proposed. Sato investigated the relationship between the rate of speech and word intelligibility under a reverberant environment [8]. In his experiments, word intelligibility at the speaking rate of 6.0 mora/s was lower than at a rate of 4.8 mora/s or less. Moreover, when the speaking rate was 3.4 mora/s, word intelligibility was the highest among that of all speaking rate under severe reverberant environment. However, it is unclear whether these tendencies can be observed under long-path echo conditions. Cui et al. reported that high familiarity words are robustly transmitted under long-path echo conditions [9]. They studied word intelligibility tests using a series of four continuous words, containing both high and low familiarity words. The results revealed that word intelligibility of high familiarity words is higher than that of low familiarity words, even when the high familiarity words overlap. This tendency is identical to the case when a high familiarity word is presented under low signal-to-noise ratio (SNR) conditions [10] or in a reverberation environment [11]. However, in a real situation, high familiarity words are not always familiar to all listeners, e.g., the name of a place for non-locals. Such words are often unfamiliar even if the other words used in the sentences are all highly familiar.

In this study, in order to investigate methods for improving intelligibility of speech under long-path echo conditions, we focused on the effect of inserting pauses between words. Since people have difficulties in understanding transmitted speech information under such severe listening conditions, they allocate more resources to perceptual processing of the incoming auditory signals. Fewer resources are available for cognitive processes when the perceptual load is higher, leading to the reduced efficiency and speed of the cognitive processing. A similar situation is observed when older people and hearing-impaired listeners listen to the sound of speech because of their degraded hearing ability. In this situation, inserting a pause between phrases [12] is an effective approach for facilitating a better understanding of speech. The insertion of pauses between phrases allow individuals more time for both perceptual processing and also for higher cognitive processes. Thus, we can expect that the insertion of pauses can also improve the understanding of speech in the presence of long-path echoes, which often cause severe listening conditions.

To investigate the effect of the insertion of pauses, the intelligibility of speech under long-path echo conditions was examined. Under such conditions, a distinguished and discrete reflection arrives with a long delay, followed by a direct sound [13]. Therefore, the long-path echo environment was simulated in this study using a direct sound and a long-path echo. Under this environment, word intelligibility tests were performed using four sequentially connected words under various pause insertion patterns and long-path echo conditions.

2. Experiment 1: Effect of a long-path echo on word intelligibility

Before investigating the effect of inserting pauses, the effect of long-path echoes on word intelligibility was analyzed in more detail than in our previous investigation [9]. Data obtained from this experiment were used as reference, and compared with the results of the experiments with pause insertion. This is discussed in the following sections.

2.1. Experimental apparatus

The experiment was conducted in a soundproof room at the Research Institute of Electrical Communications, Tohoku University. Acoustic stimuli were presented diotically using headphones (Sennheiser HDA-200) through an audio interface (Cakewalk UA-25EX) connected to a laptop computer.

2.2. Test words and experimental test conditions

In the experiment which is described later in more detail, the intelligibility of speech sounds both with and without long-path echoes was measured. Under the condition with a simulated long-path echo (henceforth long-path echo condition), the same speech sound overlapped with a specified delay, exceeding 1 s. Therefore, to investigate the effects of echoes with such a long delay, a speech sound of a suitable length was required.

When people listen to speech, particularly under severe listening conditions, they imagine the missing words by using context. The purpose of the experiment was to analyze the effect of the long-path echo on speech, and to understand this effect in detail without considering the context effect, because people rarely use this effect when they hear sentences with unfamiliar words. The experiments of the present study were similar to those in our previous study [9]; therefore, instead of actual sentences, we applied sets of four words connected sequentially as the test stimuli. By using such sequences, the effect of long-path echoes can be analyzed as a function of the position of the word. Henceforth, a series of four continuous words is called a quadruplet.

The test words were selected from a familiarity-controlled word list, called FW07 [14]. The word list consists of four word-familiarity ranks as follows: highest (7.0–5.5), second highest (5.5–4.0), second lowest (4.0–2.5), and the lowest (2.5–1.0). Each rank consists of 20 lists, and each list contains 20 words, i.e., each rank includes 400 words. All words have four moras. The words were spoken by a trained female native Japanese speaker, and recorded in a studio. Although the original sampling frequency of the recorded sound was 48 kHz, it was decreased to 16 kHz to match the sampling frequency used in Exp. 3. The quadruplets used in this experiment were composed of words with the highest familiarity ranks (7.0–5.5). The word-familiarity used in this study is recorded in the "Lexical properties of Japanese [15]," which is a word-familiarity dataset of about 88,000 word entries, derived from all word entries in a medium sized Japanese dictionary. In this dataset, word-familiarity is valued from 7 (most familiar) to 1 (most unfamiliar) for all word entries. Since only young adults were involved in judging the word-familiarity of each word in that survey, there may be individual differences of the scores among the listeners including older adults. However, it is empirically known that high familiarity words are consistently judged as highly familiar, independent of age, gender, etc [16]. Therefore, we selected words with the highest familiarity ranks.

In the experiment, the delay time to simulate a single long-path echo was treated as a parameter. Fig. 1 shows the time patterns of the presented sounds for eight conditions and different echo patterns. Here, a cluster of four symbols (circles, squares, triangles, or crosses) represents an individual word and one symbol represents each mora. Moreover, the preceding sound denotes the quadruplet which first arrives at a listening point via the shortest path. The following sound denotes the quadruplet which arrives with a delay time relative to the preceding sound, to simulate a long-path echo.

Condition 1-A consists of a preceding sound only, with no following speech sound (simulated speech sound without any long-path echoes), while conditions 1-B to 1-H consist of a preceding sound and a single following sound (speech sound with a single simulated long-path echo).

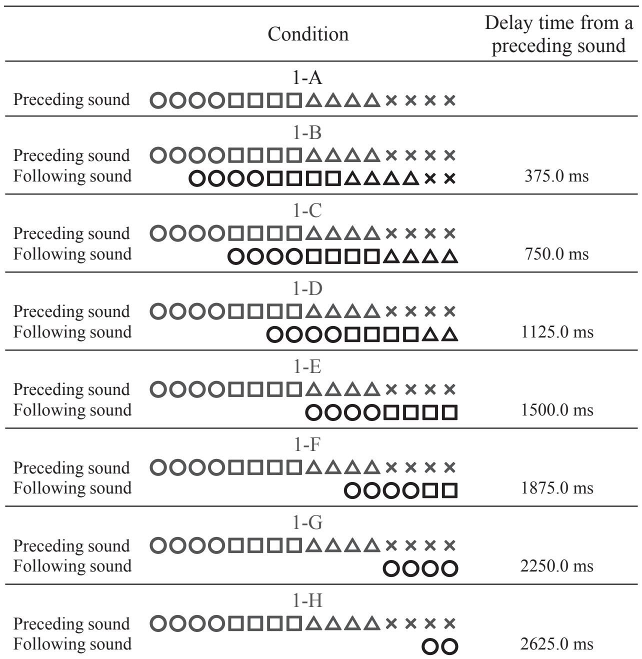


Fig. 1. Pattern diagrams of experimental conditions in Exp. 1.

As a parameter, the delay time was varied from 0 (without following sound) to 2625.0 ms in steps of 375.0 ms. The 375.0 ms step size was determined on the basis of the average word length of the 4-mora words used in the experiment. The average and standard deviations of word lengths for all words used in the experiments were 773 ms and 53 ms, respectively. The estimated word length was rounded to 750 ms. That is, the 375.0 ms step size corresponds to half the estimated word length, and this value was used as the basic unit for the delay times of the following sounds under all experimental conditions. For example, in condition 1-E, a preceding sound and a following sound with a delay time of 1500.0 ms were presented to the listeners. The preceding sound and the following sound ended at the same time as in our previous study [9].

The sound pressure levels of the spoken words were measured by 1/2 inch microphones (B&K 4192) attached to an artificial ear (B&K 4153). The A-weighted sound pressure level in terms of equivalent level (L_{Aeq}) was set to 60 dB. Speech shaped noise was added as a background noise to the speech signal to avoid the ceiling effect. The SNR was 0 dB. The speech shaped noise was added to simulate extremely severe listening condition as the reference purpose. The aim of this study was to investigate different methods of transmitting speech information under severe listening conditions. In a real-listening environment, a simulated condition could be regarded as the situation when listeners are at the edge of the service area of acoustic mass notification systems, for example. The condition is more severe than that simulated in the previous study [9].

2.3. Listeners

Twelve male and four female listeners (the average age of 22.4 years and standard deviation of 1.5) with normal hearing acuity participated in the experiment. To avoid the learning effect on the test words, each quadruplet was presented only once to each listener.

2.4. Procedure

As previously mentioned, we used FW07, which consists of 80 lists, each of which included twenty words [14]. These 80 lists were divided into four word-familiarity ranks of twenty lists, which were all used with the highest word-familiarity rank. These lists were further divided into five sets of four lists. To produce a quadruplet, the four lists in one set were assigned to the positions of the quadruplet in order (the first to the fourth). Then, one word was selected randomly from each of the four lists and the four words were connected sequentially to produce a quadruplet. All used words were discarded and this process was repeated 20 times to obtain a speech unit consisting of 20 quadruplets. This procedure was applied to the five sets. As a result, five speech units in total, consisting of 20 quadruplets each, were obtained. Because only five units were available to test eight conditions (from 1-A to 1-H), we assigned eight listeners to each condition as shown in Table 1, which shows the assignment of the speech unit × condition combinations for each listener. In this manner, each word appears only once for a specific listener. In the table, each listener is denoted by “L” followed by a number. Each listener participated in only five of the eight conditions, and listened to a different speech unit for each condition. As a result, a listener did not listen to the same word more than once. Moreover, ten listeners (2 listeners × 5 speech units) were assigned to each condition. Thus, the number of quadruplets used to test each condition was 200 (20 quadruplets × 2 listeners × 5 speech units). The word intelligibility score was calculated as the percentange of the correct answers in the 80 words (20 quadruplets). As ten listeners were assigned to each condition (Table 1), the average of the scores of these listeners represent the word intelligibility of the condition.

In one session, stimuli with the same condition were presented. Here, the 20 quadruplets were presented randomly. Listeners were asked to use a keyboard to input exactly what the speaker said as was understood. The order of the speech units was not randomized but presented from unit 1 to 5 for all listeners.

2.5. Results and discussion

The average word intelligibility scores obtained for all conditions are presented in Fig. 2. The error bars denote the standard error of the intelligibility score. By overlapping the following sound (conditions from 1-B to 1-H), the word intelligibility scores decreased compared to the case when only the direct sound was presented (condition 1-A). Analysis of variance (ANOVA) with one between-subjects factor (eight conditions) revealed a significant effect of the experimental condition ($F(7,72) = 2.80$, $p < .01$). Multiple comparisons (Ryan’s method [17], $p < .05$) revealed that the score for condition 1-A was higher than those for 1-B, 1-C, and 1-D.

Table 1
Assignment of listeners to each condition in Exp. 1.

Speech unit	Conditions							
	1-A	1-B	1-C	1-D	1-E	1-F	1-G	1-H
1	L1, L2	L3, L4	L5, L6	L7, L8	L9, L10	L11, L12	L13, L14	L15, L16
2	L15, L16	L1, L2	L3, L4	L5, L6	L7, L8	L9, L10	L11, L12	L13, L14
3	L13, L14	L15, L16	L1, L2	L3, L4	L5, L6	L7, L8	L9, L10	L11, L12
4	L11, L12	L13, L14	L15, L16	L1, L2	L3, L4	L5, L6	L7, L8	L9, L10
5	L9, L10	L11, L12	L13, L14	L15, L16	L1, L2	L3, L4	L5, L6	L7, L8

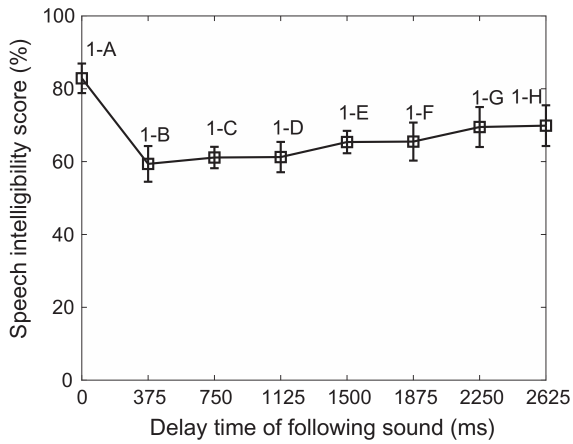


Fig. 2. Average word intelligibility scores of quadruplet in Exp. 1.

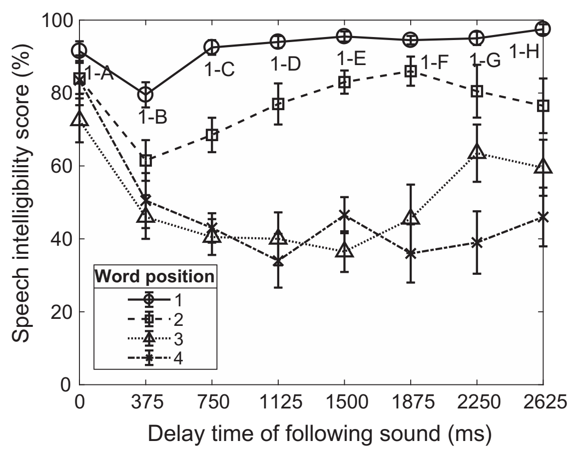


Fig. 3. Average word intelligibility scores for each word position of the quadruplet in Exp. 1.

Under certain conditions, the first word of the quadruplet can be heard with no disturbance. Therefore, the effects of the delay time on the word of each position of the quadruplet were analyzed. Fig. 3 shows word intelligibility scores for each word position as a function of the delay time. The error bars denote the standard error of the intelligibility score. Fig. 3 indicates that the effect of long-path echoes of the third and fourth words seem to greater than that of the first and second words. Analysis of variance (ANOVA) with one between-subjects factor (eight conditions) and one within-subjects factor (four positions of the word) revealed significant main effects of conditions ($F(7,72) = 2.76$, $p < .05$) and positions of the word ($F(3,216) = 233.80$, $p < .01$). The interaction between the two factors was also found to be significant ($F(21,216) = 5.72$, $p < .01$). Thus, the simple main effect of the conditions was examined, and it was found to be statistically significant at the second, the third, and the fourth words of the quadruplet (the second word: $F(7,288) = 2.11$, $p < .05$, the third word: $F(7,288) = 5.10$, $p < .01$, the fourth word: $F(7,288) = 7.38$, $p < .01$). Multiple comparisons (Ryan's method [17], $p < .05$) were performed for the intelligibility scores of the second, the third, and the fourth words to investigate the dependence of the effects of delay conditions on the position of the word in the quadruplet. In the third words, the score for condition 1-A was statistically higher than those of condition 1-B, 1-C, 1-D, 1-E, and 1-F. Moreover, the score of condition 1-G was statistically higher than that of condition 1-E. For the fourth word, the score of condition 1-A was statistically higher than those of all other conditions from 1-B to 1-H.

The aforementioned statistical analysis as well as Fig. 3 show that when a long-path echo is overlapped with the direct sound, word intelligibility decreases. This tendency is consistent with previous results [9].

3. Experiment 2: Effect of inserting pauses between 4-mora words in the quadruplet on word intelligibility

3.1. Apparatus, test words, and experimental test conditions

The apparatus and test words were identical to those used in Exp. 1.

In this experiment, both the delay time and the length of the pause were treated as parameters. Fig. 4 shows the time patterns of the presented sounds for 12 conditions for the different echo patterns and pause lengths used in the experiment. Here, a cluster of underscores represent an inserted pause. Table 2 shows the delay time and the pause length of each experimental condition as well.

Condition 2-A consists of only a preceding sound with no following speech sound (simulated speech sound without any long-path echo), while conditions 2-B to 2-L consist of a preceding sound and a single following sound (speech sound with a single simulated long-path echo). As an investigated experimental parameter, the delay time was varied from 0 (without following sound) to 937.5 ms. The pause length was set to 0 (without pause), 187.5, 375.0, and 562.5 ms. The pause length and the delay time were set to be multiples of 187.5 ms. This length was determined on the basis of the estimated mora length of the 4-mora words used in this experiment, similarly to the case of Exp. 1. The preceding sound and following sound ended at the same time, as in Exp. 1.

The sound pressure levels of the words were measured by 1/2 inch microphones (B&K 4192) attached to an artificial ear (B&K 4153). The A-weighted sound pressure level in terms of equivalent level (L_{Aeq}) was set to 60 dB. Speech shaped noise was added as a background noise to the speech signal to avoid the ceiling effect. The SNR was set to 0 dB.

3.2. Listeners

Fifteen male and nine female listeners (the average age of 21.2 years old and standard deviation of 1.2) with normal hearing acuity participated in the experiment. None of them participated in Exp. 1. To avoid the learning effect of each test speech signal, each quadruplet was presented only once to each listener.

3.3. Procedure

In this experiment, the same speech units were generated as in Exp. 1. Since only five units were available to test 12 conditions (from 2-A to 2-L), we assigned 12 listeners to each condition, as shown in Table 3, which shows the assignment of the speech unit × condition combinations for each listener. In this manner, each word appears only once for a specific listener. In the table, each listener is denoted by "L" followed by a number. Each listener participated in five of the 12 conditions only, and listened to a different speech unit for each condition. As a result, a listener did not listen to the same word more than once. Since ten listeners (2 listeners × 5 speech units) were assigned to each condition, the number of quadruplets used to test each condition was 200 (20 quadruplets × 2 listeners × 5 speech units). The word intelligibility score was calculated as a percentage of the correct answers for the 80 words (20 quadruplets). Moreover, the average scores of the ten listeners assigned to each condition (Table 3) represents the word intelligibility of the condition.

In one session, stimuli with the same condition were presented. Here, the 20 quadruplets were presented randomly. Listeners were asked to use a keyboard to input exactly what the speaker said as was understood. The order of the speech units was not randomized but presented from unit 1 to 5 for all listeners.

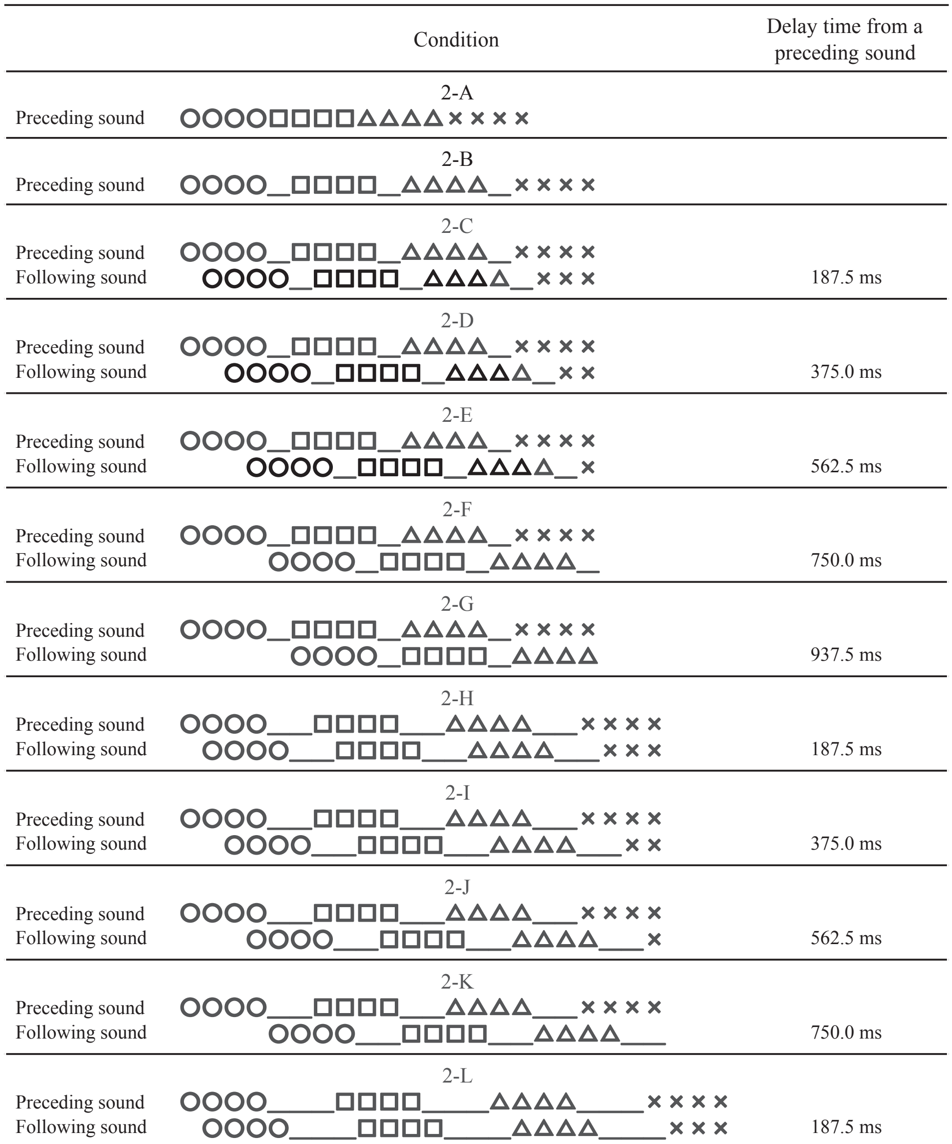


Fig. 4. Pattern diagram of experimental conditions in Exp. 2.

Table 2
Delay time and pause length of each experimental condition of Exp. 2.

Delay time (ms)	Pause length (ms)			
	0	187.5	375.0	562.5
0	2-A	2-B		
187.5		2-C	2-H	2-L
375.0		2-D	2-I	
562.5		2-E	2-J	
750.0		2-F	2-K	
937.5		2-G		

3.4. Results and discussion

The average word intelligibility score obtained for all conditions are presented in Fig. 5. The error bars denote the standard error of the intelligibility score. Except for the obtained score for condition 2-G, the scores of all the conditions are high and almost the same as the score for condition 2-A, where the following sound is not overlapped. Analysis of variance (ANOVA) with one between-subjects factor (12 pause-delay conditions) revealed a significant effect of the experimental condition ($F(11,108) = 5.29$, $p < .01$). Multiple comparisons (Ryan's method [17], $p < .05$) revealed that the score for condition 2-G was lower than those of the other conditions.

Under certain conditions, the first word of the quadruplet can be heard with no disturbance. Therefore, the effects of pause and delay time on the word of each position of the quadruplet were analyzed. Fig. 6 shows word intelligibility scores for each word position as a function of the delay time and pause length. The error bars denote the standard error of the intelligibility score. Although the observed scores under all conditions for the first word are almost identical, the scores of the other word positions are affected by the experimental conditions. The score for condition 2-G is lower than that of the other conditions in the second, third, and fourth words, as can be seen in Fig. 6(b)–(d), respectively. Analysis of variance (ANOVA) with one between-subjects factor (12 pause-delay conditions) and one within-subjects factor (four positions of the word) revealed significant main effects of the conditions ($F(11,108) = 5.29$, $p < .01$) and word positions ($F(3,324) = 5.77, p < .01$). The interaction between the two factors was also found to be significant ($F(33,324) = 5.77$, $p < .01$). Thus, a simple main effect of the pause-delay conditions was examined, and found to be statistically significant at the second, third, and fourth words of the quadruplet (the second word: $F(11,432) = 3.49$, $p < .01$, the third

Table 3
Assignment of listeners to each condition in Exp. 2.

Speech unit	Conditions											
	2-A	2-B	2-C	2-D	2-E	2-F	2-G	2-H	2-I	2-J	2-K	2-L
1	L1, L2	L3, L4	L5, L6	L7, L8	L9, L10	L11, L12	L13, L14	L15, L16	L17, L18	L19, L20	L21, L22	L23, L24
2	L23, L24	L1, L2	L3, L4	L5, L6	L7, L8	L9, L10	L11, L12	L13, L14	L15, L16	L17, L18	L19, L20	L21, L22
3	L21, L22	L23, L24	L1, L2	L3, L4	L5, L6	L7, L8	L9, L10	L11, L12	L13, L14	L15, L16	L17, L18	L19, L20
4	L19, L20	L21, L22	L23, L24	L1, L2	L3, L4	L5, L6	L7, L8	L9, L10	L11, L12	L13, L14	L15, L16	L17, L18
5	L17, L18	L19, L20	L21, L22	L23, L24	L1, L2	L3, L4	L5, L6	L7, L8	L9, L10	L11, L12	L13, L14	L15, L16

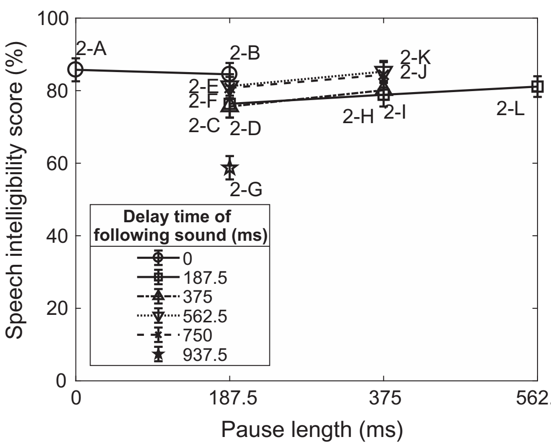
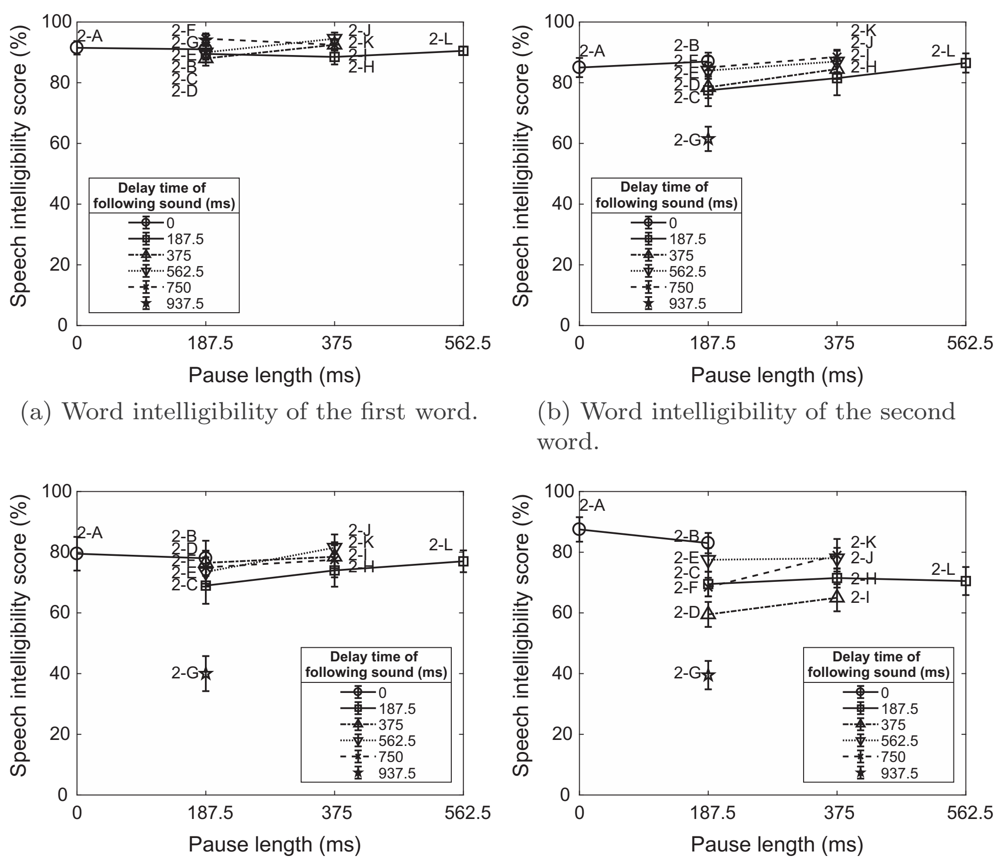


Fig. 5. Average word intelligibility score of quadruplet in Exp. 2.

word: $F(11,432) = 7.80$, $p < .01$, the fourth word: $F(11,432) = 10.22$, $p < .01$). Multiple comparisons (Ryan's method [17], $p < .05$) were performed for the intelligibility scores of the second, third, and fourth words to investigate the dependence of the effects of the pause-delay conditions on the positions of the word in the quadruplet. In the second, third, and fourth words, the score for condition 2-G was statistically lower than those of the other conditions. Moreover, in the score of the fourth word, the scores for conditions 2-A, 2-B, 2-E, 2-J, and 2-K were statistically higher than that for condition 2-D, and the score for condition 2-A was also statistically higher than those of conditions 2-F and 2-I.

The results of the statistical analysis show that there is no significant difference between the scores of the condition without long-path echo (condition 2-A) and almost all conditions with pause insertion, except for certain special cases. In particular, the score becomes statistically lower for condition 2-G compared to the other conditions in the second, third, and fourth word. This means that word intelligibility is improved to the level of the condition without long-path echo by inserting pauses between words, except for condition 2-G. The effect of inserting a pause



(a) Word intelligibility of the first word.
(b) Word intelligibility of the second word.
(c) Word intelligibility of the third word.
(d) Word intelligibility of the fourth word.

Fig. 6. Average word intelligibility scores for each word position of quadruplet in Exp. 2.

in the absence of any long-path echo cannot be observed when the score of condition 2-A is compared with that of condition 2-B, while this effect appears when a long-path echo overlaps with the direct sound, regardless of the pause length. This tendency is clearly observed at the second and the third word of the quadruplet.

4. Experiment 3: Effect of inserting a pause between 6-mora words in the quadruplet on word intelligibility

The results of Exp. 2 revealed that the insertion of pauses is effective in maintaining high speech intelligibility under various long-path echo conditions. Moreover, it is possible that the insertion of a pause of 1-mora length has a significant effect. However, this length might depend on the length of the word between pauses. To confirm that inserting a 1-mora pause is sufficient regardless of the word length, the same word intelligibility tests were performed using longer words.

4.1. Apparatus, test words, and experimental test conditions

The apparatus was identical to that used in Exp. 1. In Exp. 3, the length of each word was six moras with an accent type of LHHHLL. These words were recorded in the “Lexical properties of Japanese [15]” similarly to the 4-mora words used in Exps. 1 and 2. All words recorded in the dataset were spoken by a trained female native Japanese speaker. She was different from the speaker in Exps. 1 and 2. All words were recorded at a sampling frequency of 16 kHz. From all 6-mora words in the dataset, 457 words with familiarity of 7.0–5.0 were selected. After homonyms, words with negative meaning, and words with non-standard Japanese syllables were excluded, 336 words were selected as the test words. Some of the words are listed in Table 4.

Fig. 7 shows the time patterns of the presented sounds for seven conditions with different echo patterns and pause lengths used in the experiment. Table 5 shows the delay time and the pause length of each experimental condition. As an experimental investigated parameter, the delay time was varied from 0 (without following sound) to 1240.0 ms. The pause length was set to 0 (without pause), 155.0, and 310.0 ms. The pause length and the delay time were set to be multiples of 155.0 ms. This length was determined on the basis of the average word length of the 6-mora words used in the experiment. The average and standard deviations of word lengths for all words used in the experiments were 930.0 ms and 55.5 ms, respectively. Thus, 155.0 ms, which corresponds to the average mora length, was used as the basic unit for the pause length and the delay times of the following sounds under all experimental conditions.

As in Exp. 2, the A-weighted sound pressure level of all words was set to 60 dB and they were presented with speech-shaped noise at a SNR of 0 dB.

4.2. Listeners

Twenty-one male and seven female listeners (the average age of 21.8 years and standard deviation of 1.3) with normal hearing acuity participated in the experiment. None of them participated in Exps. 1 and 2. To avoid the learning effect of each test speech signal, each quadruplet was presented only once to each listener.

4.3. Procedure

The aforementioned selected 336 words were divided into four lists: List a, b, c, and d. Consequently, each list included 84 words. To produce a quadruplet, the four lists were assigned to the four word positions (the first to the fourth) of the quadruplet as shown in Table 6, such that each list appeared in every positions. Then, one word was selected randomly from each of the four lists and the four words were connected sequentially to produce a quadruplet. The used words were discarded and this process was repeated 84 times to obtain a speech unit consisting of 84 quadruplets.

Since only four units were available to test seven conditions (from 3-A to 3-G), we assigned each of the 28 listeners to each condition as shown in Table 7, which shows the assignment of the speech unit × condition combinations for all listeners. In this manner, each word appears only once for a specific listener. In the table, each listener is denoted by “L” followed by a number. Each listener participated in four of the seven conditions only, and listened to a different speech unit for each condition. As a result, a listener did not listen to the same word more than once. Moreover, four listeners (1 listener × 4 speech units) were assigned to each condition. Thus, the number of quadruplets used to test each condition was 84 (84 quadruplets × 1 listener × 4 speech units). The word intelligibility score was calculated as the percentage of correct answers in the 336 words (84 quadruplets). Since four listeners were assigned to each condition (Table 7), the average scores of four listeners represent the word intelligibility of the condition.

All listeners performed two sessions of the test. In the first session, half of the 84 quadruplets were presented randomly. The rest of the quadruplets were presented randomly in the second session. Listeners were asked to use a keyboard to input exactly what the speaker said as was understood.

4.4. Results and discussion

The average word intelligibility score obtained for all conditions are presented in Fig. 8. The error bars denote the standard error of the intelligibility score. While a positive effect of inserting pauses is observed in conditions 3-D and 3-F, the scores for conditions 3-E and 3-G are still low, regardless of the insertion of pauses. Analysis of variance (ANOVA) with one between-subjects factor (seven pause-delay conditions) revealed a significant effect of the experimental condition ($F(6,21) = 8.49$, $p < .01$). Multiple comparisons (Ryan’s method [17], $p < .05$) revealed that the scores for conditions 3-A and 3-F were higher than those for conditions 3-B, 3-C, 3-E, and 3-G.

Fig. 9 shows word intelligibility scores of each word position as a function of the delay time and pause length. The error bars denote the standard error of the intelligibility score. In all word positions, the observed word intelligibility scores in condition 3-F approaches the level for the condition without long-path echo (condition 3-A). The score for condition 3-D is also similar to that for condition 3-A, except for the fourth word. Analysis of variance (ANOVA) with one between-subjects factor (seven pause-delay conditions) and one within-subjects factor (four word positions) revealed significant main effects of the conditions ($F(6,21) = 8.49$, $p < .01$) and word positions ($F(3,63) = 58.91$, $p < .01$). The interaction between two factors was also found to be significant ($F(18,63) = 4.29$, $p < .01$). A simple main effect of pause-delay conditions was statistically significant at the

Table 4
Examples of 6-mora words used in the experiment.

/akirekaeru/(be completely disgusted)	/oharaibako/(be fired)	/nigirikobusi/(clenched fist)
/bideodeQki/(video deck)	/boutakatobi/(pole vault)	/wazatorasisa/(unnatural)
/masukumeroN/(cantaloupe)	/yowayowasisa/(frail)	/usirosugata/(back shot)
/hyaQkaziten/(encyclopedia)	/waraibanasi/(funny story)	/kagakuseNi/(chemical fiber)

/Q/: double consonant,/N/: syllabic nasal.

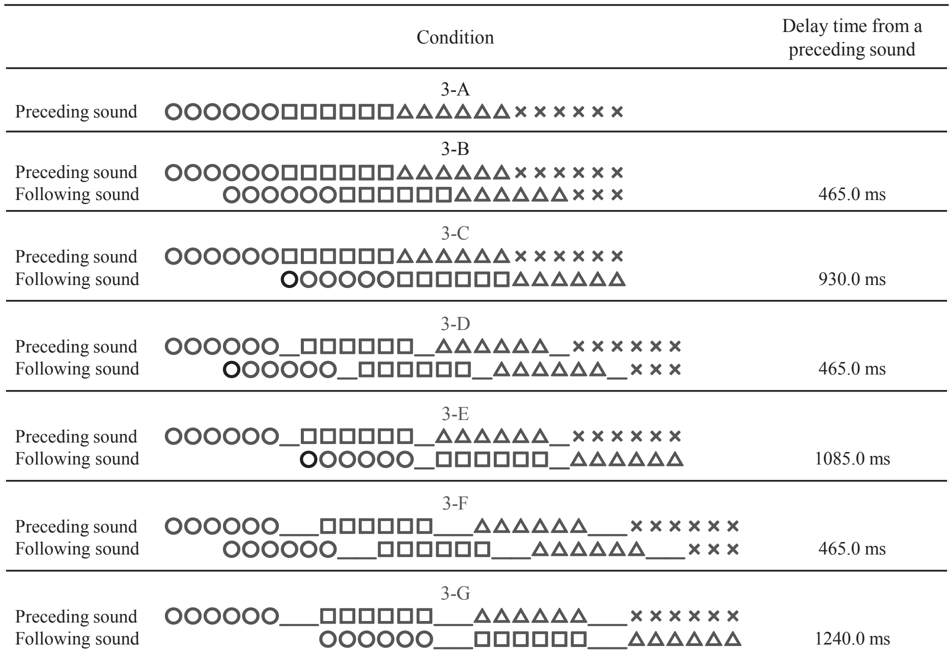


Fig. 7. Pattern diagram of experimental conditions in Exp. 3.

Table 5
Delay time and pause length of each experimental condition of Exp. 3.

Delay time (ms)	Pause length (ms)		
	0	155.0	310.0
0	3-A		
465.0	3-B	3-D	3-F
930.0	3-C		
1085.0		3-E	
1240.0			3-G

Table 6
Four generated units of quadruplets in Exp. 3.

Speech unit	Order			
	The first	The second	The third	The fourth
1	List a	List b	List c	List d
2	List d	List a	List b	List c
3	List c	List d	List a	List b
4	List b	List c	List d	List a

Table 7
Assignment of the listeners to each condition in Exp. 3.

Speech unit	Conditions						
	3-A	3-B	3-C	3-D	3-E	3-F	3-G
1	L1	L2	L3	L4	L5	L6	L7
2	L8	L9	L10	L11	L12	L13	L14
3	L15	L16	L17	L18	L19	L20	L21
4	L22	L23	L24	L25	L26	L27	L28

second, third, and fourth words of the quadruplet (the second word: $F(6,84) = 4.99$, $p < .01$, the third word: $F(6,84) = 8.12$, $p < .01$, the fourth word: $F(6,84) = 12.65$, $p < .01$). Multiple comparisons (Ryan's method [17], $p < .05$) were performed for the intelligibility scores of the second, third, and fourth words to investigate the dependence of the effects of the pause-delay conditions on the positions of the word in the quadruplet. In the score of the second word, the score of condition 3-A was statistically higher than that of condition 3-G, and the score of condition 3-F was statistically higher than those of conditions 3-B, 3-C, 3-E, and 3-G. In the score of the third word, the score of condition 3-A was statistically higher than those of conditions 3-B, 3-C, 3-E, and 3-G. The score of condition 3-D was also statistically higher than that of condition 3-G. Moreover, the score of condition 3-F was statistically higher than those of conditions 3-B, 3-C, 3-E, and 3-G. Finally, in the score of the fourth word, the score of condition 3-A was statistically higher than those of conditions 3-B, 3-C, 3-D, 3-E, and 3-G. The score of condition 3-B was statistically higher than those of conditions 3-E, or 3-G. The score of condition 3-D was also statistically higher than that of condition 3-G. Finally, the score of condition 3-F was statistically higher than those of conditions 3-C, 3-E, and 3-G.

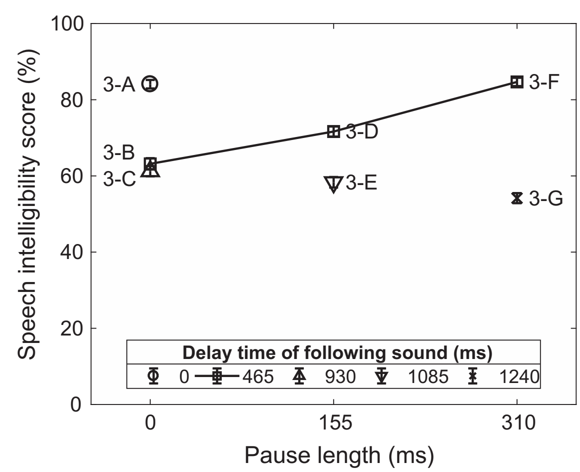
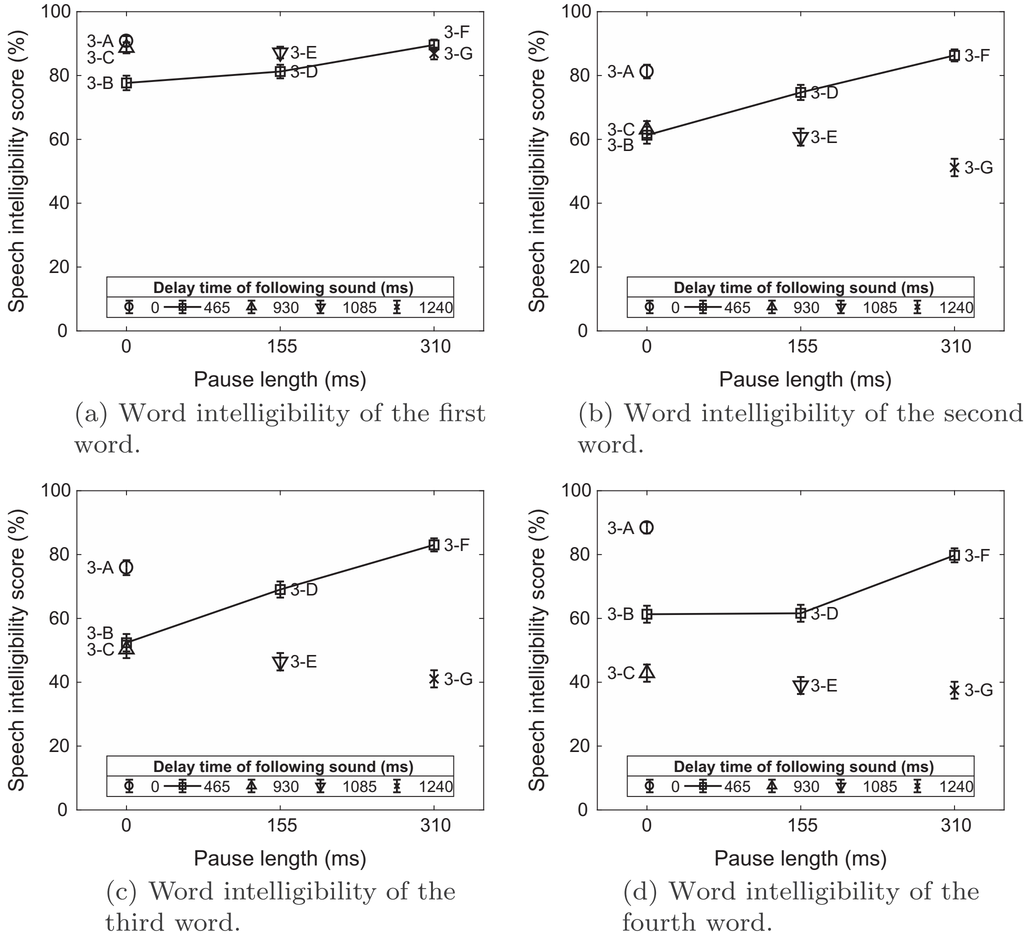


Fig. 8. Average word intelligibility score of the quadruplet.

The results of the experiment clearly show that inserting pauses effectively improves word intelligibility despite word length. Moreover, even if the length of the inserted pause is 1-mora (condition 3-D), word intelligibility is recovered at approximately the same level as for the condition without long-path echo (condition 3-A). This tendency is consistent with the results of Exp. 2.



(a) Word intelligibility of the first word.
(b) Word intelligibility of the second word.
(c) Word intelligibility of the third word.
(d) Word intelligibility of the fourth word.

Fig. 9. Average word intelligibility scores for each word position of the quadruplet in Exp. 3.

5. General discussion

5.1. The effect of inserting pauses

In this study, the effects of the insertion of pauses on word intelligibility under long-path echo conditions was studied. The results of the experiments show that word intelligibility is improved by inserting pauses between words. This effect was observed at all word positions except for the first words of the quadruplet. In this section, this effect is analyzed in more detail by comparing the results of the same long-path echo conditions of Exp. 1 and Exp. 2. In Exp. 1, the following sound overlapped with the preceding sound at a delay of 375 ms in condition 1-B. The same long-path echo condition was simulated in conditions 2-D and 2-I in Exp. 2. The only difference was the length of the inserted pause (1-B: 0 ms, 2-D:187.5 ms, 2-I: 375 ms). A similar relationship existed among condition 1-C in Exp. 1 and conditions 2-F and 2-K in Exp. 2. In this case, the delay of the following sound was 750 ms. These relationships are summarized in Fig. 10.

To analyze this effect, the score of word intelligibility at the first word of the quadruplet was excluded from the analysis, because of the case where no following sound overlapped at the first word. The scores obtained were averaged and plotted in Fig. 11. The error bars denote the standard error of the intelligibility score. By inserting a 1-mora pause or a 2-mora pause between the words in the quadruplet (conditions 2-D, 2-F, 2-I, and 2-K), the observed word intelligibility scores become more than 20 % higher than those without pauses inserted (condition 1-B and 1-C). Similar improvements are observed by comparing the word intelligibility scores for conditions 3-B, 3-D, and 3-F in Fig. 8. These results clearly indicate that inserting a pause is effective for the robust transmission of speech information under long-path echo conditions.

Here, we consider the reasons for the improved word intelligibility when pauses are inserting between words under long-path echo. One factor is the effect of simultaneous masking. By inserting pauses, people can listen to the target sound without the following sound at the part where the pause is overlapped. This means that the extent of simultaneous masking from the following sound, which is simulated as the long-path echo, decreases. As pointed out by Toida, the degradation of speech intelligibility by long-path echoes can be studied using masking [6]. According to this idea, word intelligibility is improved by inserting pauses. Moreover, there are two possibilities for individuals to listen to the target information, because the information of the following sound is completely identical to the preceding sound. Therefore, the decrease of the effect of simultaneous masking can affect an individual who is listening to the long-path echo at the part where the pause of the preceding sound is overlapped. The other factor is related to the load associated with the perceptual and cognitive processes of speech recognition. Under severe listening conditions, such as long-path echoes, people need to allocate more resources to listening to the sounds themselves. By inserting pauses between words, people can use more time not only for perceptual processing but also for higher cognitive processes. As a result, word intelligibility is improved, especially at the fourth word, which is in the most severe listening condition. This is similar to the situation when older adults listen to the sound of speech: it is known that inserting pauses facilitates speech recognition of older listeners [12]. Due to these factors, the obtained intelligibility score was higher than that obtained in the condition without echo, when pauses were inserting.

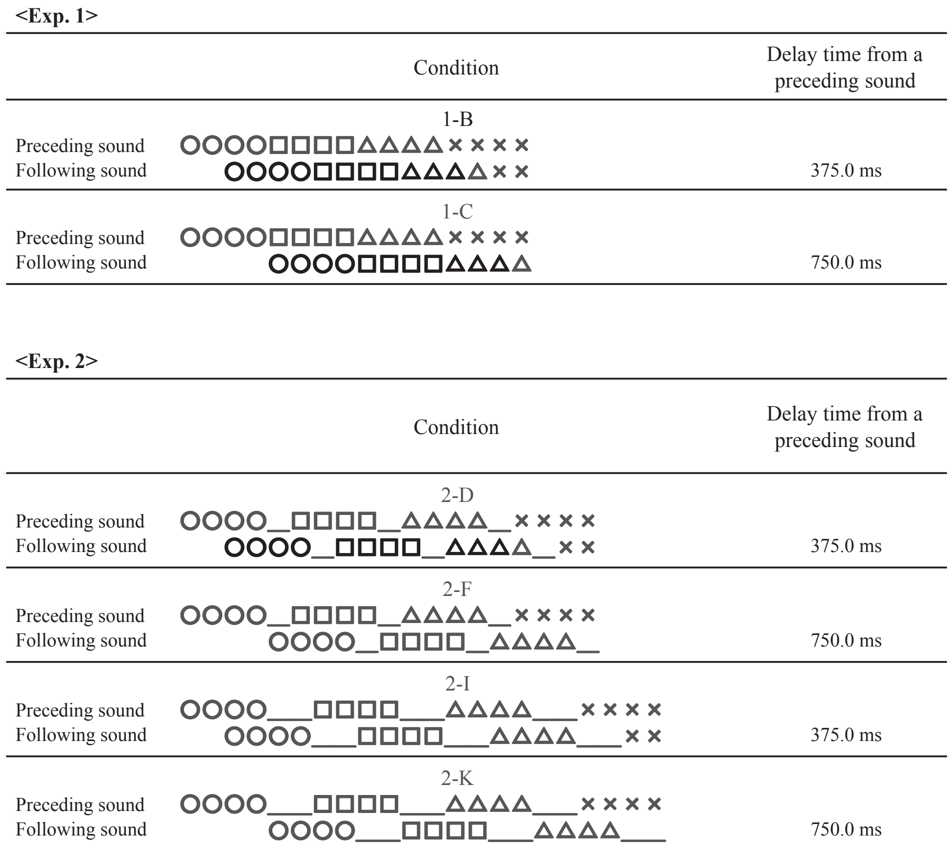


Fig. 10. Pattern diagram of the same long-path echo conditions in Exp. 1 and Exp. 2 (as shown in Figs. 1 and 4).

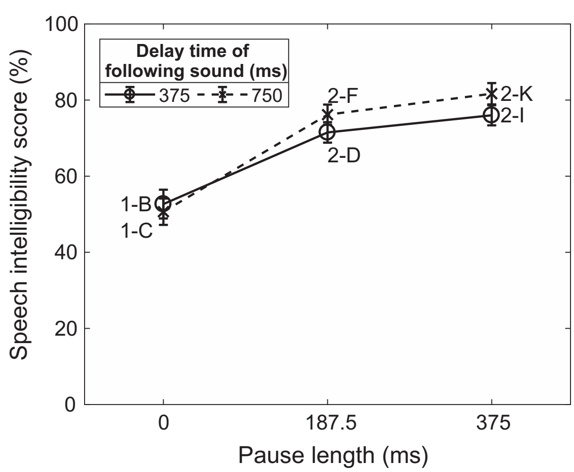


Fig. 11. Average word intelligibility score in the same long-path echo conditions in Exp. 1 and Exp. 2 with/without inserting pause.

By inserting pauses, the SNR is also increased. This increase would also be effective in improving speech intelligibility. However, the obtained improvement in word intelligibility in this study cannot be explained only by the increase of the SNR. In Fig. 11, the average word intelligibility scores under the same long-path echo conditions with/without the inserting pause is shown. As shown in Fig. 10, the difference between conditions 2-D and 2-I is only the length of the inserted pauses. In condition 2-D, 11 moras overlap each other, while eight moras can be heard without overlapping. In contrast, in condition 2-I, eight moras overlap each other, while 14 moras can be heard without overlapping. This implies that the overall SNR in condition 2-I should be higher than that in condition 2-D. However, the obtained word intelligibility score in condition 2-I is almost the same as that in condition 2-D. A similar relationship is observed between conditions 2-F and 2-K. There is almost no difference between the scores of these two conditions although the SNR should be higher in the latter. The position of the mora where the pause is overlapped also affects word intelligibility because the first mora of the word is more difficult to hear than other moras [18]. Although it is challenging to analyze the effect of each factor separately, the aforementioned effects can be expected as a result of pause insertion.

Although the positive effects of inserting pauses were observed in almost all conditions, the word intelligibility scores in condition 2-G in Exp. 2 (Fig. 5) and in conditions 3-E and 3-G in Exp. 3 (Fig. 8) were significantly lower than that for the condition without a long-path echo. Fig. 12 shows a diagram of these experimental conditions. Under these conditions, the following sounds completely overlap with the preceding sounds. This means that the amount of simultaneous masking does not decrease even when a pause is inserted. Therefore, almost no linguistic information can be obtained during these pauses. This result suggests that it is important to obtain certain linguistic information during pauses. Such situation might necessarily occur at certain areas, when all words in the sentence have equal length. However, it should be noted that the presented sentences via public-address systems usually consist of words of various lengths. Therefore, such a low intelligibility case does not happen frequently.

Analysis of each word position (Figs. 3, 6, and 9) revealed the effectiveness of inserting pauses at almost all word positions, except for the first word. However, the degree of the effectiveness was slightly different among the word positions. The effect of a long-path echo becomes more pronounced depending on the last part of the quadruplet. At the first word, the observed word intelligibility score is similarly high under all conditions. In contrast, at the third and fourth words, word intelligibility is dramatically degraded by the addition of

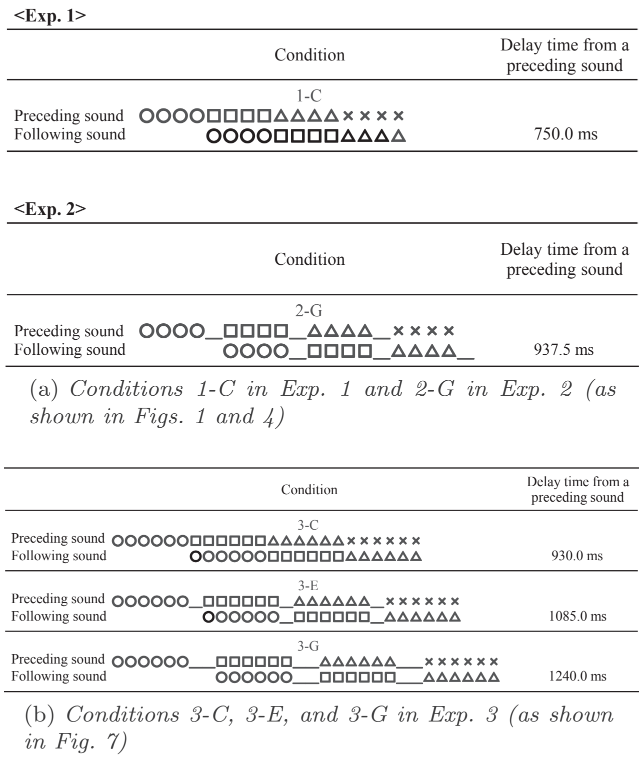


Fig. 12. Pattern diagram of the experimental conditions when all words completely overlap.

the following sound, especially in Exp. 1. The fourth word always overlapped with the following sound, while listeners can listen to the first word without any distractor. This difference causes the degradation of word intelligibility at the latter part of the quadruplet. Additionally, there is a possibility that listeners have difficulty in remembering the entire quadruplet. Sato investigated the relationship between the number of continuously presented words and word recognition scores for younger and older adults [8]. He reported that word recognition scores started to decrease when the number of continuously presented words increased over three. This decrease can be attributed to the fact that it is difficult to remember long phrases. Our results suggest that this issue of memory may enhance the decrease at the latter part of the quadruplets under severe listening conditions. Even under such severe listening conditions, at the fourth word, word intelligibility is recovered by inserting pauses between the words (Figs. 6(d) and 9(d)). By inserting a pause, listeners can allocate their perceptual and cognitive resources effectively as previous explained.

It should be noted that the last part of the following sound was cut when its preceding sound ends. Under real circumstances, the last part of the following sound is audible. Therefore, compared to a real situation, the smaller effect of the insertion of pauses observed at the fourth word in the current study might be different. If we simulate the real situation, the effect of inserting pauses can be properly observed at the fourth word, even under condition 2-D in Exp. 2 and Exp. 3. This is an important aspect of our future study.

5.2. Effective length of the pause inserted between words

In the experiments, the average of the mora-lengths of a word was used as the unit of the pause length. In the Japanese language, mora is a basic unit of language structure and Japanese speakers usually count the length of a sentence using mora. Moreover, Japanese listeners find it easier to detect segments or strings, which make up exactly one mora, than those which make up only part of a mora or span mora boundaries [19]. This evidence imply that a 1-mora pause is reasonable as the unit of the effective pause length.

As shown in Fig. 11, the score of inserting a 2-mora pause is probably slightly higher than that of inserting a 1-mora pause. However, there is no statistically significant difference in word intelligibility scores between inserting a 1-mora pause (conditions 2-D and 2-F) and a 2-mora pause (conditions 2-I and 2-K) when 4-mora words are presented as stimulus. The same tendency is obtained in Exp. 3. In Exp. 3, the score for the insertion of a 2-mora pause is apparently higher than that of inserting a 1-mora pause (Fig. 8), although there is no statistically significant difference between the scores of inserting a 2-mora pause (condition 3-F) and a 1-mora pause (condition 3-D). These results imply that at least 1-mora length is required to improve word intelligibility under long-path echo condition.

By inserting a 1-mora pause, audible moras are increase. Due to this, compensation of the missing mora from the audible mora is facilitated. In the experiments, only high familiarity words were used as test words. Word familiarity is strongly related to lexical information. That is, the higher the word familiarity is, the richer the lexical information. This implies that listeners can easily compensate for missing mora from other audible moras in the experiments. Indeed, the results of our previous study revealed that word intelligibility of high familiarity words is higher than that of low familiarity words under long-path echo conditions [9]. Therefore, high word intelligibility score is observed even when a 1-mora pause is inserted.

A 1-mora length is also closely related to the sufficient pause length required for older adults to understand speech information. Tanaka et al. reported that a 200 ms pause between phrases was sufficient for older adults to enhance speech recognition [12]. The inserted 1-mora pause in Exp. 2 was 187.5 ms, which is almost same as the sufficient pause length reported by Tanaka et al. 200 ms is also within the rage of the length of the inserted pause used in Exp. 3. Although the effective unit of pause length should be investigated in future studies, these results suggest that by inserting at least a 1-mora pause, speech information can be transmitted more robustly under long-path echo conditions than without inserting these pauses.

As suggested by the results, 1-mora is confirmed to be an effective length of the pause to transmit speech information robustly under long-path echo conditions. However, the most suitable length of the pause for effectively transmitting speech information under such conditions is still an open question. Although a longer pause may be better according to the results of Exp. 3, the length of the pause has to be set as short as possible to transmit emergency information quickly to the listeners. Considering these factors, it is important to determine the optimum pause length in the future studies.

6. Conclusions

Under long-path echo conditions, reflected long-path echo sounds from the surrounding environment cause significant degradation of the intelligibility of presented speech. To transmit speech information under such conditions, it is important to analyze the effect of long-path echoes on speech intelligibility, and to overcome this effect by controlling the characteristics of the presented sounds.

In this study, we considered the effects of inserting pauses between the words in a sentence. The results from word intelligibility tests using quadruplets show that word intelligibility under long-path echo conditions can be improved to a level approaching those in conditions without long-path echoes, by inserting pauses between the words. The result shows that speech communication robust to long-path echoes can be achieved, using speech signals with inter-word pauses. The results in the present study also indicate that inter-word pauses of a length of approximately 1-mora is required.

Acknowledgements

A part of this study was supported as a research and development project (Research and development strengthening disaster-proof properties of information and communication networks) by the 3rd supplementary budget of the Ministry of Internal Affairs and Communications, Japan, the A3 Foresight Program for "Ultra-realistic acoustic interactive communication on next-generation Internet," and JSPS KAKENHI Grant No. JP16KT0100.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.apacoust.2018.01.020.

References

[1] ISO 9613-1: Acoustics – Attenuation of sound during propagation outdoors – Part 1: Calculation of the absorption of sound by the atmosphere. Geneva (CH): Standard, International Organization for Standardization; 1993.
[2] Bolt RH, MacDonald AD. Theory of speech masking by reverberation. J Acoust Soc Am 1949;21(6):577–80. http://dx.doi.org/10.1121/1.1906551.
[3] Morimoto M, Sato H, Kobayashi M. Listening difficulty as a subjective measure for evaluation of speech transmission performance in public spaces. J Acoust Soc Am 2004;116(3):1607–13. http://dx.doi.org/10.1121/1.1775276.
[4] Toida Y. Sentence intelligibility of reinforced speech under echoic and disturbing noise conditions (in Japanese). Trans Architect Inst Japan 1984;346:112–23.
[5] Toida Y. Speech intelligibility in open air sound fields (in Japanese). J Acoust Soc Japan 1987;43(7):519–25.
[6] Toida Y. Characteristics and prediction of speech intelligibility in reinforced sound fields disturbed by echoes with a long delay time (in Japanese). J Archit, Plan Environ Eng 1997;62(492):1–8.
[7] Sato H, Morimoto M, Sakamoto S, Suzuki Y. The effect of long-path echo on intelligibility of the disaster radio system (in Japanese). In: Proceedings ot the Spring Meeting on Acoustics Society of Japan, no. 2-Q-33; 2013. p. 949–52.
[8] Sato H. Effect of number of keyword in a chunk of spoken messages and speaking rate on word recognition performance and listening difficulty under noisy and reverberant sound fields (in Japanese). In: Trans. Tech. Comm. Psychol. Physiol. Acoust., Acoust. Soc. Jpn., vol. 38; 2008. p. 1–6.
[9] Cui Z, Sakamoto S, Morimoto M, Suzuki Y, Sato H. Effect of word familiarity on word intelligibility of four continuous word under long-path echo conditions. Appl Acoust 2017;124:30–7. http://dx.doi.org/10.1016/j.apacoust.2017.02.001.
[10] Amano S, Sakamoto S, Kondo T, Suzuki Y. Development of familiarity-controlled word lists 2003 (fw03) to assess spoken word intelligibility in Japanese. Speech Commun 2009;51(1):76–82. http://dx.doi.org/10.1016/j.specom.2008.07.002.
[11] Sato H, Sato H, Yoshino H, Suzuki Y, Amano S, Kondo T, et al. Effects of word familiarity and hearing loss with aging on word intelligibility in noise and reverberation (in Japanese). J Acoust Soc Jpn 2002;58:346–54.
[12] Tanaka A, Sakamoto S, Suzuki Y. Effects of pause duration and speech rate on sentence intelligibility in younger and older adult listeners. Acoust Sci Technol 2011;32(6):264–7. http://dx.doi.org/10.1250/ast.32.264.
[13] Takashima K, Aoki M, Tsuru H, Mitsueda T, Koike H, Sato H, et al. Measurement and prediction of sound propagation impulse response in outdoor (in Japanese). In: Proceedings of the Spring Meeting on Acoustical Society of Japan, no. 2-Q-32; 2013. p. 947–8.
[14] Speech Resources Consortium < http://research.nii.ac.jp/src/en/FW07.html > .
[15] Amano S, Kondo T. Lexical properties of Japanese (in Japanese), no. 1. Sanseido (Tokyo); 1999.
[16] Sato H, Sato H, Yoshino H, Suzuki Y, Amano S, Kondo T, et al. Effects of word familiarity and hearing loss with aging on word intelligibility in noise and reverberation (in Japanese). J Acoust Soc Jpn 2002;58(6):346–54.
[17] Howell DC. Statistical methods for psychology fifth edition, 5th ed. Duxbury (CA); 2002.
[18] Sakamoto S, Amano S, Suzuki Y, Kondo T, Ozawa K, Sone T. The effect of familiarity on mora identification in word intelligibility tests (in Japanese). J Acoust Soc Jpn 2004;60:3351–7.
[19] Warner N, Arai T. Japanese mora-timing: a review. Phonetica 2001;58(1–2):1–25. http://dx.doi.org/10.1159/000028486.