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).