Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Intelligibility of speech spoken in noise and reverberation

Nao Hodoshima (1), Takayuki Arai (2) and Kiyohiro Kurisu (3)

(1) Department of Information Media Technology, Tokai University, 2-3-23 Takanawa Minato-ku, Tokyo 108-8619, Japan
(2) Department of Information and Communication Sciences, Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo, 102-8554, Japan
(3) TOA Corporation, 2-1 Takamatsu-cho, Takarazuka-shi, Hyogo, 665-0043, Japan

PACS: 43.70.MN, 43.72.DV, 43.71.ES, 43.55.HY, 43.55.JZ

ABSTRACT

When we speak in an environment with noise, we often modify our speech production, and speech spoken in noise is generally more intelligible than speech produced in a quiet environment, which is known as the Lombard effect. Our goal is to provide intelligible speech announcements in noisy and/or reverberant public spaces, such as train stations. Thus, the present paper examines whether speech spoken in reverberation as well as in noise improves speech intelligibility than speech spoken in quiet. We recorded words in a carrier sentence produced by two native speakers of Japanese in quiet, white noise, and reverberation (reverberation time of 3.6 s). We added white noise to the recorded speech at a signal-to-noise ratio of -2 dB or convolved it using one of two impulse responses (reverberation times of 3.6 and 2.6 s). Then we conducted listening tests with the speech sounds for 32 young native speakers of Japanese. The results show that the speech spoken not only in noise but also in reverberation is more intelligible than speech spoken in quiet. The intelligible reverberation-induced speech is observed regardless of whether speakers and listeners were in the same or different reverberant situations. The results further show that modification of the pre-target phrase contributes more to the improvements of speech intelligibility in reverberation than modification of the target word.

INTRODUCTION

In noisy and reverberant public spaces (e.g., train stations), speech announcements are sometimes difficult to hear. This is a general tendency to elderly people, people with hearing impairments, and non-native listeners. To overcome this problem, several solutions have been used, mainly in the fields of architectural acoustics (e.g., wall treatments) and electroacoustics (e.g., loudspeaker arrangements). Recently a signal processing solution has been proposed, which would be implemented in a sound reinforcement systems [1, 2]. These solutions require construction/renovation of public spaces and electroacoustic systems.

It may also be possible to improve speech intelligibility in public spaces by focusing on the speech announcement itself, or in other words, the nature of speech production. This is because we often modify our speech to make it robust against noise and/or reverberation, which is known as the Lombard effect in a noisy environment [e.g., 3]. Several studies have reported that the acoustic characteristics of speech produced in noise are modified in temporal and spectral domains, compared to those of speech spoken in quiet: word amplitude is higher, word duration is longer, and pitch, the first and second formant frequencies (F1 and F2), is higher in Lombard speech than in speech produced in quiet [4, 5]. We have observed similar modifications in reverberation, although the changes are not exactly the same between noisy and reverberant environments [6].

It has also been shown that Lombard speech is more intelligible than speech uttered in quiet. Speech produced in noise has higher word identification scores than speech uttered in quiet at a -5 to -15 dB signal-to-noise ratio (SNR) of babble noise and white noise [4, 5].

It has not been well clarified, however, whether speech produced in reverberation is intelligible in the same way as reported for speech in noise [3-5]. Noise and reverberation have different temporal and spectral masking patterns on speech. For example, simultaneous masking has been observed in noise, while overlap-masking (i.e., the overlap of the energy of a preceding phonemes on the following ones) occurs in reverberation [7]. Thus, there is no correlation between the ongoing speech and the noise a speaker hears, while in reverberation an ongoing speech signal and the reverberant one a speaker hears have a correlation. A similar mechanism in the latter case is delayed auditory feedback (DAF). The intensity and pitch of speech produced under DAF increase [8], which have also been reported in Lombard speech [3-5]. However, speech under DAF also shows deterioration of fluency, and this often leads to a decrease in intelligibility [8].

Our goal is to provide intelligible speech announcements in noisy and/or reverberant public spaces. In the present work, as a step towards achieving this goal, we focused on speech production and examined whether speech spoken in noise or reverberation is more intelligible than speech spoken in quiet when the speech sounds are heard in noise or reverberation.

For the reverberation environment, we further examined how speech intelligibility changed when the recording and listening conditions (e.g., reverberation time) were the same or different. We also examined which part of a sentence produced in reverberation contributed to improvements in speech intelligibility (either a target word or a pre-target phrase). This paper first describes the recording for obtaining speech samples produced in quiet, noise, and reverberation. It then describes listening tests we performed in noise or reverberation using young participants and discusses the results.

RECORDING

Speakers

Two native speakers of Japanese (one male and one female, 20 and 22 years old) served as speakers. They reported normal hearing and no articulation disorders.

Speech samples

Speech samples consisted of 36 target words in a carrier sentence. The target words were four morae and selected from a database of familiarity-controlled Japanese word lists [9]. The familiarity of the target words used in the current study were between 2.5 and 4.0 on a 7-point scale (1, most unfamiliar; 7, most familiar) [9].

We used three speaking conditions in the recording (Table 1): quiet (Q), noise (N), and reverberation (R1). White noise was used as “N”. An impulse response for “R1” we used was recorded in a church. It had reverberation time of 3.6 s at an average of octave bands from 125-4000 Hz.

Recording settings

Figure 1 shows the recording settings. The recording was conducted in a sound-treated room. Sounds were recorded on a computer through a microphone (SHURE, KSM109), amplifier (PreSonus, DIGIMAX FS), and digital audio interface (RME, Fireface 800). Noise was presented to the speakers over headphones (SENNHEISER, HDA200; dynamic, closed circumaural type) through the digital audio interface. As reverberant sounds, a sound picked up by the microphone was convolved with the impulse response and presented to the speakers over the headphones through the digital audio interface. Noise addition and an impulse response convolution were processed in real time by using Adobe Audition 3.0. The recorded sounds contain neither noise nor reverberation.

A practice session was held to familiarize the speakers with the recording procedure. During the practice session, the playback level of noise and reverberant sounds was set to -22 dB (A-weighted sound pressure level was used in this paper) relative to the speaking level of the speakers at their ears, and this level was maintained throughout the recording.

Each speaker conducted 108 trials (36 sentences x 3 speaking conditions). In each trial, the speakers were instructed to read a sentence aloud twice with a short interval between readings. The sentence was displayed on a computer monitor, which was about 1.0 m from the speakers. They were instructed to imagine that their speech is being broadcasted to a public space with room acoustics as they hear and read the sentence as clearly as possible. The recording started with Q and then preceded to N and R. The order of sentences in each speaking condition was randomized across the speakers.

Table 1. Speaking conditions.

Speaking conditions	Sounds presented over headphones
quiet (Q)	
noise (N)	white noise
reverberation (R1)	reverberant speakers' utterance

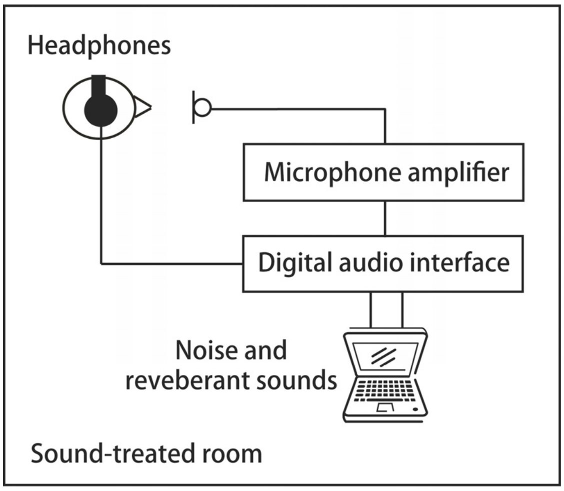

Fig. 1. Recording settings.

LISTENING TEST

Participants

Thirty-two native speakers of Japanese (four males and 28 females; average age of 23 years) participated in the listening test. All of them had normal hearing; air-conduction thresholds were less than 25 dB HL from 125 to 8 kHz for both ears.

Stimuli

As we have seen in the previous section, the recorded speech (Figure 2, top) consisted of the target word embedded in the carrier sentence that is divided into a pre-target phrase and a post-target phrase. In reverberation, a pre-target phrase affects target intelligibility due to the reverberant masking. Although the speakers read aloud the same carrier sentence in the recording, each utterance differs acoustically. Thus, acoustically the same pre-target phrase and the post-target phrase were used for all target words within the same speaking condition in order to control the effect of the pre-target phrase on the target intelligibility. Stimuli were made by splicing the pre-target phrase, the target word and the post-target phrase together (Figure 2, bottom). The combination of the carrier sentence and target word was Q (carrier sentence) and Q (target word), N and N, R1 and R1, and Q and R1 (four spliced conditions). The combination was made within a same speaker. The intensity ratio (in dB) of the carrier sentence relative to the target word was normalized within the spliced condition according to one for the selected carrier sentence.

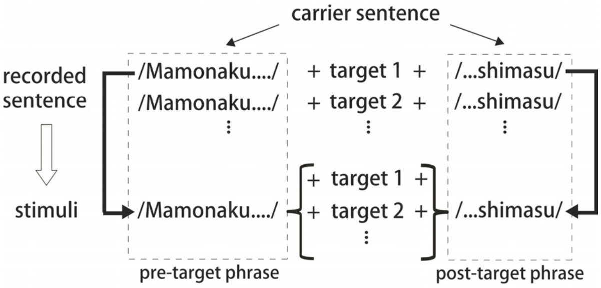

Fig. 2. Stimuli preparation.

Table 2. Listening conditions.

Listening condition	Recording	Listening test
Q_N	Q	N
N_N	N	
Q_R1	Q	R1
R1_R1	R1	
QR1_R1	carrier: Q target: R1	
Q_R2	Q	R2
R1_R2	R1	
QR1_R2	carrier: Q target: R1	

As for the listening conditions, either white noise was added or the impulse response was convolved with the spliced sentences. The conditions used in the listening test are shown in Table 2, where Q_N means the recording was done in quiet and the listening test was conducted in noise, N_N that both the recording and listening test were conducted in noise, and so on. The SNR was -2 dB for the listening test. Two impulse responses were used in the listening test: the one used in the recording (R1; reverberation time of 3.6 s) and an impulse response that was a multiplication of R1 by an exponential decay, which gave reverberation time of 2.6 s (R2). Overall intensity (in dB) was scaled across the listening conditions and speakers.

Procedure

The listening test was conducted in an anechoic room. The stimuli were presented to the participants diotically over headphones (STAX, SR-303; electrostatic, open circumaural type) through a digital audio interface (Onkyo, MA-500U) connected to a computer. Two practice trials were held to familiarize the participants with the procedure beforehand. The playback level was adjusted to the participants' comfort level. In each trial, a stimulus was presented once, and the participants were instructed to write down what they heard as the target words on an answer sheet. When the participants clicked a computer screen, the next trial was presented. For each participant, 32 stimuli (8 listening conditions x 4 spliced conditions) were randomly presented. All 34 words including two words for the practice session were different for each participant. Combinations of the target words and the listening condition were counter-balanced across the participants.

RESULTS AND DISCUSSION

Figure 3 shows the mean percent correct of mora in each listening condition for each speaker (S1 and S2) as well as for the average of the speakers. According to the design of the counter-balance used in the listening test, we carried out separate statistical analyses for the main effect of two speakers and eight listening conditions. Separate Tukey's multiple comparisons were also carried out for N, R1 and R2.

There was no main effect of the speakers, indicating that the overall correct rates were the same between them. The main effect of the listening conditions was significant (*p*<0.01). Further separate results in noise and reverberation are as follows.

N_N had significantly higher mora identification scores than Q_N (*p*=0.003). This is consistent with the other Lombard speech studies [3-5], meaning that noise-induced speech is more intelligible than speech produced in quiet when it's heard in noise.

For the reverberant conditions, R1_R1 had significantly higher mora identification scores than Q_R1 (*p*=0.019) and QR1_R1 (*p*=0.001). Further, R1_R2 had significantly higher mora identification scores than Q_R2 (*p*=0.027) and QR1_R2 (*p*=0.018). These results indicate that speech spoken not only in noise but also in reverberation is more intelligible than in a quiet environment, although the speech masking mechanism is different between noise and reverberation.

The results in both reverberant conditions indicate that reverberation-induced speech (R1_R1 and R2_R2) is more intelligible than speech spoken in quiet (Q_R1 and Q_R2), regardless of whether speakers and listeners are in the same or different reverberant situations. This implies that speech announcements become more intelligible when a speaker hears reverberant utterances during a recording. The result further implies that, for making speech announcements, it is not necessary to set up a reverberant condition that is equal to one at a public space in order to yield higher speech intelligibility in reverberation.

There was no significant difference between Q_R1 and QR1_R1 or between Q_R2 and QR1_R2. This means that speech intelligibility was not significantly improved when the target word was changed from Q to R1 (or R2). On the other hand, R1_R1 and R1_R2 had significantly higher correct rates than QR1_R1 and QR1_R2, respectively. This means that speech intelligibility was significantly improved when the pre-target word was changed from Q to R1 (or R2). The acoustic characteristics of reverberation-induced speech are modified in temporal and spectral domains [6], and one of them is longer silences between segments in reverberation-induced speech than speech produced in quiet. Thus, the reverberation masking of a pre-target phrase to a target word is more decreased in the reverberant-induced speech compared to the speech produced in quiet when it's heard in reverberation. This indicates that modification of the pre-target phrase contributes more to the improvements of speech intelligibility in reverberation than modification of the target word. In other words, not only a target word itself but also a pre-target phrase should be taken into account for making speech announcements in reverberation.

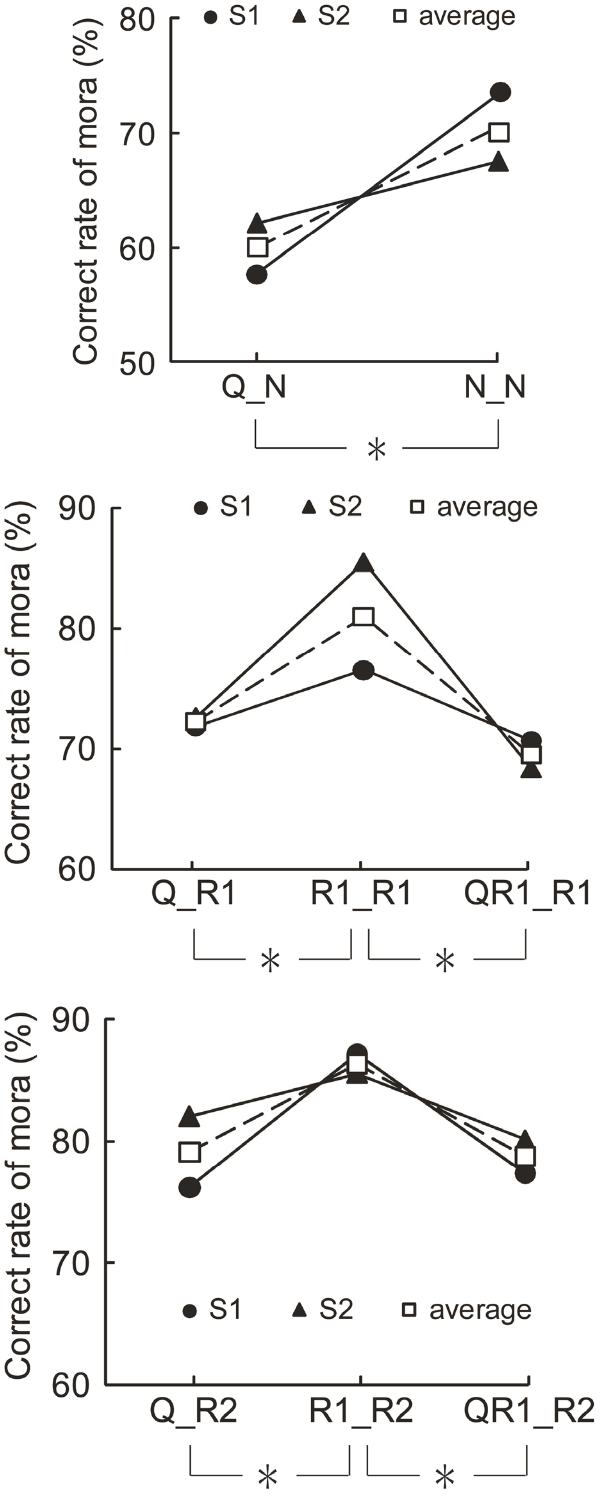

Fig. 2. Mean percent correct of mora in each listening condition [top, listening test in noise; middle, listening test in reverberation (R1); bottom, listening test in reverberation (R2)] for each speaker (circles and triangles) as well as for the average of speakers (squares). An asterisk shows a significant difference at p<0.01.

CONCLUSIONS

We recorded speech sounds in quiet, white noise, and reverberation and carried out listening tests that presented the recorded speech to young participants in noise and reverberant environments. The results showed that noise-induced speech, as reported in [3-5], but also reverberant-induced speech is more intelligible than speech spoken in quiet, when they are heard in noise or reverberation. We also showed that the reverberation-induced speech is more intelligible, regardless of whether speakers and listeners are in the same or different reverberant situations. This means that the reverberation condition for making speech announcements and that for a public space do not have to be the same, which make it easier to improve speech intelligibility in public spaces in practice. The results further imply that, in order to make intelligible speech announcements in reverberation, speakers need to pay attention not only to a main word (e.g., a station name) but also to a preceding phrase.

Future study will carry out acoustic analyses of the noise/reverberant-induced speech to find the acoustic characteristics that correlate to higher intelligibility of those speech sounds. It will be also interesting to examine which part of preceding phrase (e.g., pause, intensity and duration) contributes to speech intelligibility of the noise/reverberant-induced speech. A possible application of the current findings would be in instructing speakers how to effectively transmit messages to their audiences in public spaces where relatively higher speech intelligibility is required. Another possible approach is to develop a speech synthesis system that is suitable for noisy and reverberant public spaces.

ACKNOWLEDGMENTS

This work was partially supported by Grant-in-Aid for Young Scientists (B) from MEXT (21700203) and by Sophia University Open Research Center from MEXT. We appreciate Hideki Tachibana, Kanako Ueno and Sakae Yokoyama for providing the impulse response data.

REFERENCES

1 T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto and T. Kitamura, “Effects on suppressing steady-state portions of speech on intelligibility in reverberant environments,” *Acoust. Sci. Tech*., 23(4), 229-232 (2002)
2 N. Hodoshima, T. Arai, A. Kusumoto and K. Kinoshita, “Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments,” *J. Acoust. Soc. Am*., 119(6), 4055-4064 (2006)
3 H. Lane H and B. Tranel, “The Lombard sign and the role of hearing in speech,” *J. Speech Hear. Res*., 14, 677-709 (1971)
4 W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow and M. A. Stokes, “Effects of noise on speech production: Acoustics and perceptual analysis,” *J. Acoust. Soc. Am*., 84, 917-928 (1988)
5 J.-C. Junqua, “The Lombard reflex and its role on human listeners and automatic speech recognizers,” *J. Acoust. Soc. Am*., 93, 510-524 (1993)
6 N. Hodoshima, T. Arai and K. Kurisu, “Speaker variabilities of speech in noise and reverberation,” *IEICE Technical Report*, SP2009-69, 43-48 (2009) (in Japanese)
7 A. K. Nabelek, T. R. Letowski and F. M. Tucker, “Reverberant overlap- and self-masking in consonant identification,” *J. Acoust. Soc. Am*., 86, 1259-1265 (1989)
8 A. J. Yates, “Delayed auditory feedback,” *Psychological Bulletin*, 60, 213-232 (1963)
9 S. Amano, T. Kondo, S. Sakamoto and Y. Suzuki, “Familiarity-controlled word lists 2003 (FW03),” *The Speech Resources Consortium, National Institute of Informatics in Japan* (2006)