DOI: 10.61782/fa.2025.0860

FORUM ACUSTICUM
EURONOISE 2025

The effect of congruence of virtual image and reverberation-induced speech on speech intelligibility

Nao Hodoshima[1*] **Kazuhiko Hamamoto**[2] **Mitsunori Mizumachi**[3]
Ken-Ichi Sakakibara[4]
[1] Department of Information and Telecommunication Engineering, Tokai University, Japan
[2] Department of Information Media Technology, Tokai University, Japan
[3] Department of Electrical and Electronic Engineering, Kyushu Institute of Technology, Japan
[4] Department of Communication Disorders, Health Sciences University of Hokkaido, Japan

ABSTRACT[*]

Public address announcements under reverberation are generally unintelligible, and reverberation-induced speech (reverberant version of Lombard speech) may offer a solution. Since the audio-visual information people receive in public spaces is not always consistent, this study investigates the role of audio-visual congruency in the intelligibility of reverberation-induced speech. Two virtual classroom images were created by Unity: Smaller room simulated a space with a reverberation time (RT)=1.5 s, and Larger room simulated a space with RT=12 s. A young adult recorded sentences under quiet condition (Q), congruent reverberation (R1: Larger room and RT=12 s), and incongruent reverberation (R2: Larger room and RT=1.5 s and R3: Smaller room and RT=12 s). Under R1-R3, the reverberant speech was fed back via headphones, and the images were displayed via a head-mounted display (HMD). Twenty young adults carried out word identification tests wearing headphones and HMD under reverberation. The results showed that R1 was significantly more intelligible under RT=12 s than Q and R3. However, no significant difference was found under RT=1.5 s between Q and R2, suggesting that audio, rather than visual information, contributes to the intelligibility of reverberation-induced speech when RT is relatively long.

Keywords: *Lombard effect, reverberation, speech intelligibility, audio-visual, CG*

1. INTRODUCTION

Public address (PA) announcements are essential for protecting our lives in emergencies. However, compared with quiet environments, noise and reverberation degrade speech intelligibility, especially for older adults (OAs) than for young adults (YAs) [1]. Therefore, PA announcements need to be sufficiently intelligible.

People modify their vocal efforts in noisy environments to make themselves more intelligible, a phenomenon known as the Lombard effect [2]. This effect, which is characterized by increased word intensity, duration, fundamental frequency, and formant frequencies, has been shown to enhance speech intelligibility for various talker and listener groups in noisy environments [2-6]. When talkers speak in the presence of reverberation (reverberation-induced speech), the acoustic characteristics and speech intelligibility of that speech are increased [7,8] similarly to those observed in the Lombard effect, despite the different masking patterns of the noise and reverberations.

It is widely known that visual information improves speech intelligibility, especially in adverse listening environments or for low-predictability sentences [e.g., 6,9]. YAs and OAs received significant intelligibility benefits in speech reception threshold from auditory and auditory-visual Lombard speech in a speech-shaped noise. Compared with a case with no cues, preceding congruent and incongruent

[*]***Corresponding author***: *hodoshima@tokai.ac.jp*

contextual cues have been shown to increase and decrease word intelligibility, respectively, in a stationary noise [9].
In emergencies in public spaces, available text information (e.g., smartphones or electronic bulletin boards) may not coincide with PA announcements, and this incongruency may degrade the intelligibility of PA announcements. Previous studies [10,11] investigated the effects of normal/urgent speech and preceding congruent/incongruent texts on word intelligibility for YAs and OAs in noisy and reverberant environments. The results demonstrated that the word correct rate was significantly higher for urgent speech than for normal speech and congruent text than incongruent or no text. The urgent speech benefit was the same regardless of the preceding text condition.
Displaying virtual images by head-mounted display (HMD) gives us a more immersive experience and study under more realistic environments in audio-visual research. However, the effect of virtual images on speech perception remains unclear. Localization accuracy of the target talker significantly decreased with an increasing number of interfering talkers (four or more) and reverberation time; however, audio-visual congruency did not significantly affect the result when the virtual room was modeled using Unity and presented HMD [12]. The degree of reverberation did not significantly differ among audio only, congruent and incongruent audio-visual stimuli where the virtual room is rendered from 3-D geometric models to panoramic images using Blender and presented by HMD [13]. As far as the authors know, the effect of audio-virtual visual conditions on speech intelligibility, especially on Lombard speech and reverberation-induced speech, has not been investigated.
This study investigates how audio-virtual visual congruency/incongruency affects speech intelligibility of reverberation-induced speech in reverberant environments. Recording and listening tests were conducted by displaying congruent and incongruent CG images by HMD. Two extreme reverberant conditions (1.5 s and 12 s) were used as a preliminary study.

2. LISTENING TEST

2.1 Participants

Twenty native speakers of Japanese (age range, 20–23 years) participated in this study. All participants reported no history of hearing aid use and normal or corrected-to-normal vision.

2.2 Stimuli

Two virtual visual rooms are used: the smaller room and the larger room. The smaller room is 10 m in length × 10 m in width × 2 m in height for simulating a classroom for 40 people. The larger room was 100 times larger than the small room. The wall surface textures were fabric for the smaller room and concrete for the larger room. The rooms were modeled using Unity and rendered on HMD. The audio conditions are reverberant speech with two different reverberation times (RTs). Four conditions were used in recording: quiet Q (no reverberation speech and visual stimuli fed back), congruent R1 (the larger room and RT=12 s), and incongruent (R2: the larger room and RT=1.5 s, R3: the smaller room and RT =12 s).
The speech material was a target word embedded in the carrier sentence, “People in the classroom immediately evacuate to [target word]. (English translation)” Forty-four target words consisting of four morae (the smallest Japanese phonological unit) were selected from a database of familiarity-controlled lists [14]. The word familiarity used in the present study was between 2.5 and 4.0 on a seven-point scale (1: least familiar to 7: most familiar) to prevent participants from using context and semantic cues.
The recording was made on a PC through a microphone (Shure KSM141; condenser, cardioid) and a digital audio interface (Roland OCTA-CAPTURE UA-1010) in a sound-treated room at a sampling frequency of 44100 Hz and a bit-depth of 24 bits. The talker was a female native speaker of Japanese aged 24 years with no history of hearing or voice disorders. Under Q, the talker recorded the speech material without wearing headphones and HMD. Under R1-R3, the talker wore headphones (Sennheiser HDA200; dynamic, closed circumaural type) and HMD (Meta Quest 3), and visual stimuli were projected on the HMD. The carrier sentence and the target words were displayed in the middle of the visual stimuli. Talker utterances were convolved by an impulse response (RT was either 1.5 s or 12 s), and reverberant sounds were fed to the talker through headphones. The playback level was set to ‑10 dB relative to the speaking level at the talker’s ears. The recording was controlled by Adobe Audition.
After recording, to control the effect of overlap-masking [15] on the target words, a carrier sentence was chosen for each condition, and the target words were embedded in the carrier sentence for each condition. Finally, the concatenated speech sounds were convolved with either of the impulse responses using MATLAB. Table 1 shows five listening conditions. The overall intensity of the stimuli was normalized across all conditions. The total number of

Table 1. Listening conditions.

Condition	Visual stimuli	Audio recording	RT(s)
A		Quiet	1.5
B		Quiet	12
C (congruent)	Larger room	Reverberation-induced	12
D (incongruent)	Larger room	Reverberation-induced	1.5
E (incongruent)	Smaller room	Reverberation-induced	12

stimuli was 244 (five conditions × 40 sentences + four sentences for the practice trial).

2.3 Procedures

The listening test was carried out in a sound-treated room. Audio stimuli were presented to each participant diotically over the headphones through a digital audio interface (Presonus Studio 26c) connected to a PC. Visual stimuli were projected on the HMD. Four practice trials were held with the target words not used in the main trial. The playback level was adjusted to each participant's comfort listening level. In each trial, a stimulus was presented once, and the participants were instructed to record the target word as they heard to the microphone via the audio interface. For each participant, 40 stimuli (five conditions × eight sentences) were presented randomly. The combination of the target word and condition was counter-balanced across the participants.

3. RESULTS

Fig. 1 shows the percentage of correct mora answers for each condition. One-way analysis of variance was carried out with conditions as the repeated variable and the correct rate as the dependent variable. The main effect was significant ($F(4, 76) = 512.93$, $p < 0.01$). A post hoc test revealed that Condition A has significantly higher correct rate than Condition B ($p < 0.01$), Condition C ($p < 0.01$), and Condition E ($p < 0.01$), Condition C has significantly higher correct rate than Condition B ($p = 0.0385$) and Condition E ($p = 0.0299$), and Condition D has significantly higher correct rate than Condition B ($p < 0.01$), Condition C ($p < 0.01$), and Condition E ($p < 0.01$).

A significantly higher correct rate in Condition C than in Condition B indicates that reverberation-induced speech had higher speech intelligibility than speech spoken in quiet,

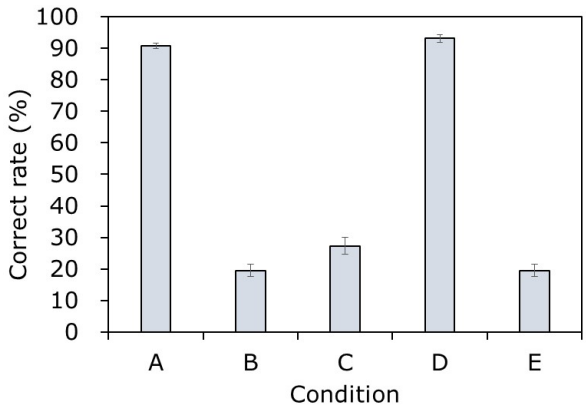

Figure 1. Mean correct rate and standard error of target word.

as in the previous research [7,8]. However, it is unclear how visual information further increased the intelligibility of reverberation-induced speech compared with audio-only stimuli. That Condition C had a significantly higher correct rate than Condition E indicated that the congruent audio-visual speech improves speech intelligibility than incongruent audio-visual speech under the same RT, which is consistent with the previous research [10,11] while the previous research evaluated the urgent speech.

The lack of significant difference in the correct rates between Conditions A and D and between Conditions B and E suggests that the speech intelligibility of incongruent reverberant-induced speech did not differ from that of speech spoken in a quiet environment under the same RT. This contradicts previous research [10,11], where visual stimuli were texts on a PC screen. The difference in visual information between this study and previous ones [10,11] might have affected the correct rates.

The results revealed that speech intelligibility was not affected by virtual visual room information, along with the localization accuracy of the target talker in noise and the perceived degree of reverberation [12,13]. However, it is unclear if speech production is affected by virtual visual room information (i.e., vocal effort might increase in the larger room). Further acoustic analysis will reveal how virtual visual information affects speech production and speech perception.

4. CONCLUSIONS

This study investigated the effect of audio- virtual visual congruency/inconsistency on speech intelligibility of reverberation-induced speech in reverberant environments.

Quiet speech and reverberation-induced speech were recorded, whereas, in the latter condition, congruent/incongruent reverberant speech and virtual rooms were presented to a talker by headphones and HMD. Word identification in sentence tests was conducted using quiet speech and congruent/incongruent reverberation-induced speech and virtual rooms in reverberant environments. Results showed that 1) Congruent reverberation-induced speech had a significantly higher correct rate than speech spoken in quiet and incongruent reverberation-induced speech under the same RT, 2) Speech intelligibility of incongruent reverberation-induced speech did not significantly differ from that of speech spoken in a quiet environment under the same RT, and 3) Speech intelligibility was not affected by virtual visual room information. Further acoustic analysis and testing of different reverberant conditions will reveal how virtual visual information affects speech production and perception of reverberation-induced speech, contributing to appropriate evacuation PA announcements.

5. ACKNOWLEDGMENTS

This work was supported by a Grant-in-Aid for Scientific Research B from the Japan Society for the Promotion of Science (#21H01596). We are grateful to Hideki Tachibana, Kanako Ueno, and Sakae Yokoyama for providing the impulse response data and to Koushin Sugawara and Haruki Shioda for carrying out the experiment.

6. REFERENCES

[1] A. K. Nablek and P. K. Robinson, “Monaural and binaural speech perception in reverberation for listeners of various ages”, J. Acoust. Soc. Am., 71(5), 1242-1248, 1982.

[2] H. Lane and B. Tranel, “The Lombard sign and the role of hearing in speech”, J. Speech Hear. Res., 14(4), 677-709, 1971.

[3] W. Van 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(3), 917-928, 1988.

[4] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise", J. Acoust. Soc. Am., 124(5), 3261-3275, 2008.

[5] M. Cooke and M. L. G. Lecumberri, “The intelligibility of Lombard speech for non-native listeners”, J. Acoust. Soc. Am., 132(2), 1120–1129, 2012.

[6] M. F. Fitzpatrick, J. Kim, and C. Davis, “Auditory and auditory-visual Lombard speech perception by younger and older adults”, Proc. International Conference on Auditory-Visual Speech Processing, 105–110, 2013.

[7] N. Hodoshima, T. Arai, and K. Kurisu, "Intelligibility of speech spoken in noise and reverberation", Proc. International Congress on Acoustics (paper ID: 663), 2010.

[8] N. Hodoshima, T. Arai, and K. Kurisu, "Intelligibility of speech spoken in noise/reverberation for older adults in reverberant environments", Proc. Interspeech (paper ID: P6a.06), 2012.

[9] A. A. Zekveld, M. Rudner, I. S. Johnsrude, J. M. Festen, J. H. M. Beek, and J. van Rönnberg, “The influence of semantically related and unrelated text cues on the intelligibility of sentences in noise”, Ear Hear., 32, E16-E25, 2011.

[10] N. Hodoshima, “Effects of urgent speech and congruent/incongruent text on speech intelligibility in noise and reverberation”, Proc. Interspeech. 3113-3117, 2019.

[11] N. Hodoshima, “Effects of urgent speech and congruent/incongruent text on speech intelligibility for older adults in the presence of noise and reverberation”, Speech Commun., 134, 12-19, 2021.

[12] A. Ahrens and K. D. Lund, “Auditory spatial analysis in reverberant multi-talker environments with congruent and incongruent audio-visual room information”, J. Acoust. Soc. Am. 152 (3), 1586-1594, 2022.

[13] M. Schutte, S. D. Ewert, and L. Wiegrebe, “The percept of reverberation is not affected by visual room impression in virtual environments”, J. Acoust. Soc. Am., 145 (3), EL229–EL235, 2019.

[14] 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.

[15] A. K. Nabelek, T. R. Letowski, and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification", J. Acoust. Soc. Am., 86(4), 1259-1265, 1989.