INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Effects of urgent speech and congruent/incongruent text on speech intelligibility in noise and reverberation

Nao Hodoshima

Department of Information Media Technology, Tokai University, Japan
hodoshima@tokai-u.jp

Abstract

Public-address (PA) announcements are widely used, but noise and reverberation can render them unintelligible. Furthermore, in an emergency, textual information available to smartphone users or displayed on electronic bulletin boards may not coincide with PA announcements, and this mismatch may degrade the intelligibility of PA announcements. This study investigated how speech spoken in a normal/urgent style and preceding congruent/incongruent textual information affected word intelligibility and perceived urgency in noisy and reverberant environments. The results obtained from 18 participants showed that the word correct rate (WCR) was significantly higher for urgently spoken speech than for normal speech, and for congruent text than for incongruent/no text. However, there was no speaking style-text interaction, indicating that the improvement in WCR provided by urgent speech over normal speech was the same regardless of the preceding text condition. This suggests that listeners rely more on visual information when speech intelligibility is poor. The results for perceived urgency also showed that the congruent condition was rated "evacuate now", while the incongruent condition was rated "wait and see". These results suggest that simple combinations of speaking style and textual information decrease the intelligibility of emergency PA announcements, and audio-visual incongruence must be considered.

Index Terms: speech intelligibility, urgent speech, incongruent text, reverberation, noise

1. Introduction

Public address (PA) announcements are sometimes difficult to understand because of noise and reverberation in public spaces. Additionally, noise and reverberation generally degrade speech intelligibility to a greater extent for older adults and non-native listeners than for young native listeners [1, 2]. Because the world's population is rapidly aging (e.g., the elderly population rate in Japan was 27.7% in 2018 [3]), this must be considered when delivering PA announcements, especially in emergency situations.

One way to increase the intelligibility of PA announcements is for people to modify the way they speak according to the surrounding acoustic environment (e.g., the Lombard effect [4]). Speech spoken in noisy environments has been shown to yield higher word identification scores than that spoken in quiet environments when heard in noisy environments [4–6]. When talkers speak in an environment with reverberation, the intelligibility of that speech was found to increase in a manner similar to the cases observed as the Lombard effect [7, 8].

Speaking slowly is another way to increase the intelligibility of PA announcements, especially in reverberant environments, because it reduces overlap-masking [9]. Time-compressed speech (with a time compression rate of 40%) resulted in lower intelligibility for older adults in noisy or reverberant environments than for young adults [10]. Speech slowed by a time-delay technique in which speech sounds sent from loudspeakers are delayed based on the distance between adjacent loudspeakers and the speed of sound resulted in decreased perceived listening difficulty relative to that for normal-speed speech [11]. However, speaking too slowly may not be appropriate for emergency PA announcements because it may be difficult for the public to perceive the level of urgency, which may lead to delayed evacuation.

Urgently spoken words have a greater perceived urgency than normal speech in quiet environments [12], and urgent speech has been shown to have a greater perceived urgency and higher speech intelligibility than normal speech spoken at the same speed in noisy and reverberant environments [13]. Speech spoken in such an urgent style is characterized by a higher fundamental frequency, a broader fundamental frequency range, and a higher amplitude, and this style has been shown to yield higher perceived urgency than speech spoken in a normal style [12, 13]. In another study, it was demonstrated that urgently spoken words yielded faster response times than normally spoken words [14], suggesting that urgent speech may be more effective for emergency evacuation than normal speech.

It is widely known that prior knowledge improves speech intelligibility, especially in adverse listening environments. In a previous study, when a written word was presented before a vocoded word, the clarity rating of the vocoded word was enhanced with matching text and was reduced with mismatching text, relative to the case with no written information [15]. The same result was observed for speech perception in stationary noise: matching and mismatching three-word contextual cues respectively increased and decreased word perception relative to the case with no word cues [16]. Speech cues instead of text also showed the same effect. Preceding words increased the identification rate of the target word in comparison with the case with no preceding words [17]. In an emergency, textual information available to smartphone users or displayed on electronic bulletin boards may not coincide with PA announcements, and this mismatch may degrade the intelligibility of PA announcements. To the best of the author's knowledge, the relationship between congruent or incongruent text and the intelligibility of urgent speech in noise and reverberation has not been previously investigated.

The goal of this study was to develop a method of making PA announcements more intelligible in noisy and reverberant environments by modifying the speech itself rather than

http://dx.doi.org/10.21437/Interspeech.2019-1902

implementing architectural acoustic and/or electroacoustic solutions. The current study investigated whether the inclusion of congruent or incongruent text preceding urgent speech affects its intelligibility and perceived urgency in noisy and reverberant environments.

2. Listening test

2.1. Participants

Eighteen young native speakers of Japanese (mean age: 23 years) participated in this study. All reported normal hearing and normal or corrected to normal vision.

2.2. Stimuli

Speech materials consisted of target words embedded in the carrier sentence, "A fire has broken out. Evacuate to [target word]." Sixty-two target words of four morae (a phonological syllable-like unit in Japanese) were selected from a database of familiarity-controlled Japanese word lists [18] with word familiarity between 1.0 and 2.5 on a seven-point scale (1: least familiar, 7: most familiar) to avoid the participants using context and semantic cues. The target words used here were the same as those used in [16]. Semantic information is known to change the perceived urgency [12]; however, this study was concerned with the direct effect (bottom-up cues) of urgent speech on speech intelligibility in noisy and reverberant environments.

Speech materials were recorded on a computer through a condenser microphone (SHURE KSM141) and a digital audio interface (TASCAM US-144MKII) at a 44100 Hz sampling rate in a sound-treated room. The talker was a 24-year-old female native speaker of Japanese and was the same as the speaker used in [16]. She received voice training at a voice acting school for two years and has experience as an amateur voice cast in a play. After the speech materials were recorded, one carrier sentence was chosen for each speaking condition, and the target words with 100-ms pauses before and after were embedded in the carrier sentence in each speaking condition. This was done to control the effect of overlap-masking on the target words. The intensity ratio of the carrier sentence relative to the target word was normalized.

Table 1 shows the six considered experimental conditions, consisting of two different speaking conditions (normal and urgent) and three different preceding text conditions (no text, congruent text, incongruent text). In the urgent condition, the talker was instructed to imagine that she was making a PA announcement that a fire had broken out in a train station and to warn passengers with urgency about the emergency. In the normal condition, the talker was instructed to speak as she speaks normally. In the congruent and incongruent text conditions, written text was presented on a computer monitor before the speech stimulus was presented. In the congruent text condition, the written text was the same as the target word. In the incongruent text condition, the written text was different from the target word.

All audio stimuli were combined with a babble noise (a mixture of four utterances from two male talkers from a speech database [19]) with a signal-to-noise ratio of 8 dB and then convolved with an impulse response (reverberation time of 2.0 s for octave bands from 125 to 4000 Hz) using MATLAB to simulate an average listening environment in a subway station installed with sound-reflective walls. The overall intensity of the stimuli was normalized. Then the text and audio stimuli were combined using video editing software (GOM mix pro) in MP4 format. Figure 1 shows the timeline of the stimuli presentation.

Table 1: *Experimental conditions.*

Condition	Details
1	Normal speech
2	Urgent speech
3	Congruent text + normal speech
4	Congruent text + urgent speech
5	Incongruent text + normal speech
6	Incongruent text + urgent speech

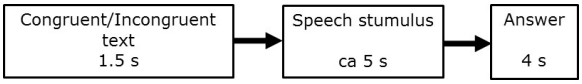

Figure 1: *The timeline of stimuli presentation*

The total number of stimuli sets was 362 (6 conditions × 60 sentences + 2 sentences for a practice session). The six conditions were fully crossed with 10 trials in each condition. Trials were randomly ordered during each block. The target words assigned to each condition were randomized over the participants, and each target word was presented once per block.

2.3. Procedures

A listening test was carried out for each participant in the sound-treated room. Two practice trials were held to familiarize the participants with the experimental procedure. The playback level was adjusted to each participant's comfort level.

In each trial, the congruent or incongruent text was first presented for 1.5 s on a computer monitor. Written text was composed of black characters presented on a white background. (In the no text condition, only the speech stimulus was presented.) After the text was removed from the monitor, the speech stimulus was presented once to participants diotically through headphones (STAX SR-303; electrostatic, open circumaural type) through a digital audio interface (TASCAM US-144MKII) connected to the computer. The participants were instructed that the text may or may not help them understand the speech sounds. After the speech stimulus was presented, the participants were given 4 s to write down what they heard as the target word on their answer sheets. Then the participants were asked to rate the impression of the stimulus. The rating scales were (1) "evacuate now" or "wait and see," and (2) ease of remembering the evacuation announcement on a five-point scale (1: difficult to remember, 5: easy to remember). Instead of directly asking participants about the perceived urgency [12, 14], it was investigated how the preceding text and urgent speech separately affected the participants' impressions by using adjective pairs. For each participant, 60 stimuli (6 conditions × 10 sentences) were presented randomly.

3. Results

Figure 2 shows the mean correct rate of the target words. Statistical analysis was carried out using an analysis of variance (ANOVA) with speaking condition (normal and urgent) and preceding text (no text, congruent text, and incongruent text) as

the independent variables and the correct rate of the target words as the dependent variable.

The main effect of the speaking condition was significant ($F(1, 17) = 7.149$, $p = 0.016$), indicating that urgent speech was significantly more intelligible than normal speech. The main effect of the preceding text condition was also significant ($F(2, 34) = 86.582$, $p < 0.01$), and post-hoc multiple comparisons revealed that the identification rate of words preceded by congruent text was significantly higher than that of speech preceded by either incongruent text or no text ($p < 0.01$).

Figure 3 shows the number of participants rating each stimulus case as "evacuate now" or "wait and see." A Friedman test was conducted among repeated measures and showed significance ($\chi^2 = 27.033$, $p < 0.01$). Post-hoc multiple comparisons revealed significant differences between normal and congruent normal ($p = 0.046$), normal and congruent urgent ($p = 0.008$), urgent and congruent urgent ($p = 0.046$), urgent and incongruent normal ($p = 0.008$), urgent and incongruent urgent ($p = 0.005$), congruent normal and incongruent normal ($p = 0.011$), congruent normal and incongruent urgent ($p = 0.020$), congruent urgent and incongruent normal ($p = 0.001$), and congruent urgent and incongruent urgent ($p = 0.001$).

Figure 4 shows the ratings of the participants' impressions of how easy they considered it to remember the evacuation announcements on a five-point scale. A Friedman test was conducted among repeated measures and showed significance ($\chi^2 = 58.327$, $p < 0.01$). Post-hoc multiple comparisons revealed significant differences between normal and urgent ($p = 0.01$), normal and congruent normal ($p = 0.002$), normal and congruent urgent ($p = 0.003$), normal and incongruent normal ($p = 0.010$), and urgent and incongruent urgent ($p = 0.012$).

4. Discussion

As has been reported previously [12, 13], urgent speech was significantly more intelligible and showed increased perceived urgency (more responses for "evacuate now") than normal speech. Speech preceded by congruent text showed significantly higher speech intelligibility than speech preceded by incongruent text or no preceding text. This is partly consistent with previous studies using vocoded speech [15] and listening tests in noisy environments [16]. In that previous study, incongruent text was associated with a reduced word correct rate in comparison with the no text condition [16], whereas in this study, no difference was found in the word correct rates for the no text and incongruent text conditions. This disagreement may be due to the number of preceding words and availability of semantic information (the previous study [16] used three-word cues, whereas this study used single-word cues). These results indicate that not only speaking style but also preceding text affect speech intelligibility and contextual information helps listeners compensate for degraded acoustic information [17]. Several papers have claimed that this effect of written text on signal detection bias is due to top-down processing; such claims have been based on listening tests investigating the effect of changing the presentation timing of text relative to the onset of a speech stimulus on the clarity rating of vocoded words [15] and on experiments involving tracking the activity of the auditory cortex by functional magnetic resonance imaging (fMRI) during the perception of vocoded speech [20]. However, simulations with a phonemic decision-making model have indicated that top-down processing is not necessary because lexical information is combined with phonological information only at a late decision stage when phonological judgment is

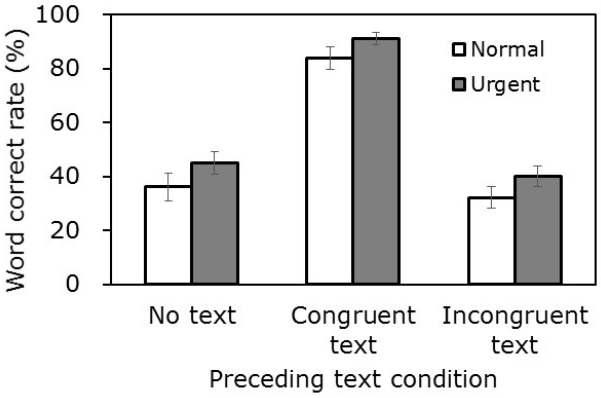

Figure 2: *Mean word correct rate and standard error of target words in speech spoken normally or urgently preceded by congruent, incongruent, or no text.*

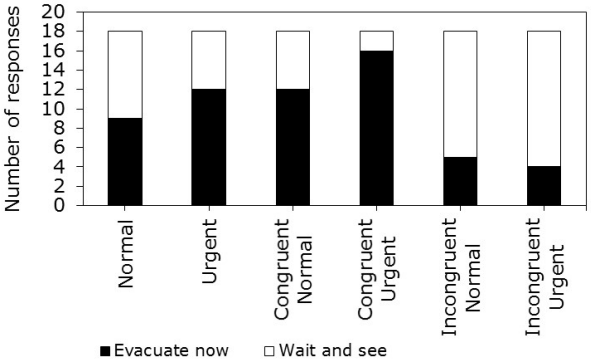

Figure 3: *Number of ratings of "evacuate now" or "wait and see" for speech spoken normally or urgently preceded by congruent, incongruent, or no text.*

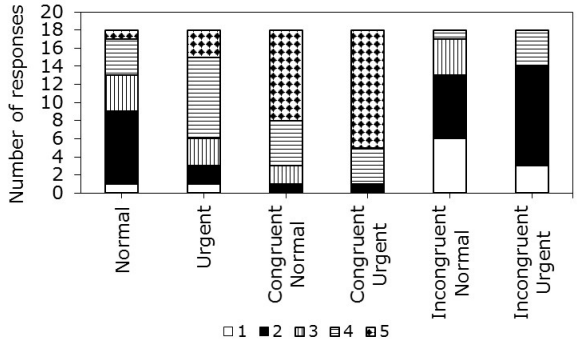

Figure 4: *Number of ratings of 1 to 5 reflecting ease of remembering evacuation announcements (1: difficult to remember, 5: easy to remember) for speech spoken normally or urgently preceded by congruent, incongruent, or no text.*

formed [21].

Interestingly, there was no significant interaction between the speaking style and preceding text conditions, indicating that the improvement in speech intelligibility provided by urgent speech over normal speech was the same regardless of the preceding text condition. Previous findings that the effect of the preceding text on word intelligibility is not affected by changing

the signal-to-noise ratio [16] and the results of the present study support the conclusion that listeners rely more on visual speech information when auditory intelligibility is poor. This is further supported by similar results from a previous study [22], in which face and lip information were provided with speech stimuli simultaneously.

In the impression rating results, congruent normal and congruent urgent speech were more often rated as "evacuate now" than normal and urgent speech, respectively, whereas incongruent speech was more often rated as "wait and see" than speech without text and congruent speech. Furthermore, congruent normal, congruent urgent, and urgent speech were more often rated as easy to remember than normal speech, whereas incongruent normal and incongruent urgent speech were more commonly rated as difficult to remember than normal and urgent speech, respectively. From the intelligibility results, the preceding text condition has more of an effect than the speaking style on the impression rating. Because there was no significant difference in speech intelligibility between speech without preceding text and speech preceded by incongruent text, incongruent speech was more commonly rated as "wait and see" and difficult to remember than speech without preceding text. This indicates that PA speech accompanied by incongruent announcement text may delay emergency evacuation because people may be confused about which information is correct.

The inclusion of a preceding sound (an ocean wave-like sound) has been shown to improve the intelligibility of urgent speech in comparison with normal speech with no preceding sound [13], and urgent speech has been shown to yield faster response times than normal speech [14]. These results suggest that a preceding sound may have an alerting effect and aid in the understanding of PA announcements, especially for situations in which the spoken announcement and text are incongruent. It would be interesting to investigate how preceding text and sounds affect the intelligibility of urgent speech and perceived urgency in noisy and reverberant environments, especially in incongruent audio–visual situations.

Although semantic information is known to affect the perceived urgency rate [12] and it has been shown that female talkers tend to yield higher ratings of perceived urgency than male talkers [23], context and semantic cues were not addressed in this study, and only a female talker was used to create the speech stimuli. In future research, the scope may be expanded to explore the effects of talker gender and semantic information on the intelligibility and perceived urgency of speech in noisy and reverberant environments.

5. Conclusions

The present study investigated how combinations of speaking styles (normal and urgent) and preceding text (no text, congruent text, and incongruent text) affect word correct rate and perceived urgency in noisy and reverberant environments. The results obtained from 18 participants showed that urgent speech was significantly more intelligible than normal speech, which is consistent with results obtained in previous studies. Speech preceded by congruent text was significantly more intelligible than speech preceded by incongruent text or no text. There was no significant interaction between speaking style and preceding text condition, showing that the improvement in speech intelligibility by urgent speech over normal speech was the same regardless of preceding text type, supporting the conclusion that listeners rely more on visual speech information when auditory intelligibility is poor.

The results also showed that congruent normal and congruent urgent speech were more often rated as "evacuate now" than normal and urgent speech, respectively, whereas incongruent speech was more rated as "wait and see" than speech without text and congruent speech. Furthermore, congruent normal, congruent urgent, and urgent speech were more often rated as easy to remember than normal speech, whereas incongruent normal and incongruent urgent speech were more often rated as difficult to remember than normal and urgent speech, respectively. Because there was no significant difference in speech intelligibility between speech without preceding text and speech preceded by incongruent text, incongruent speech was more often rated as "wait and see" and difficult to remember than speech without preceding text. This indicates that incongruence between a spoken PA and any accompanying text may delay emergency evacuation because people may be confused about which information is correct.

Future research will involve the inclusion of semantic information and context cues as well as an increased number of talkers and diverse listeners such as older adults, who are more affected by noise and reverberation than young adults. This study revealed that simple combinations of speaking style and textual information may decrease the intelligibility of emergency PA announcements, and audio–visual incongruence must be considered. Further research is needed to determine appropriate combinations of speaking styles and alerting sounds and text with the aim of further increasing the intelligibility of emergency PA announcements.

6. Acknowledgements

The author is grateful to the talker and listeners who participated in this study, and to Kazumasa Okonogi of Tokai University for conducting listening tests.

7. 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*., vol. 71, pp. 1242-1248, 1982.

[2] A. K. Nabelek and A. M. Donahue, "Perception of consonants in reverberation by native and non-native listeners", *J. Acoust. Soc. Am*., vol. 75, no. 2, pp. 632-634, 1984.

[3] The Japanese Cabinet Office, "Annual report on the aging society", 2018.

[4] H. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech", *J. Speech Hear. Res.*, vol. 14, pp. 677-709, 1971.

[5] 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*., vol. 84, pp. 917-928, 1988.

[6] J. C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers", *J. Acoust. Soc. Am*., vol. 93, pp. 510-524, 1993.

[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. K. Nabelek, T. R. Letowski and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification", *J. Acoust. Soc. Am*., vol. 86, pp. 1259-1265, 1989.

[10] S. Gordon-Salant and P. J. Fitzgibbons, “Recognition of multiply degraded speech by young and elderly listeners”, *J. Speech Hear. Res*., vol. 38, pp. 1150-1156, 1995.
[11] S. Yokoyama, S. Sakamoto, H. Tachibana and S. Tazawa, “Study on the application of time-delay technique to public address system in a tunnel”, *Proc. Inter-noise*, 2005.
[12] E. Hellier, J. Edworthy, B. Weedon, K. Walters, A. Adams, “The perceived urgency of speech warnings: Semantics versus acoustics” *J. Human Factors*, vol. 44, no. 1, pp. 1-17, 2002.
[13] N. Hodoshima, “Effects of urgent speech and preceding sounds on speech intelligibility in noisy and reverberant environments.” Proc. Interspeech, pp. 1696-1699, 2016.
[14] J. K. Ljungberg and F. Parmentier, “The Impact of intonation and valence on objective and subjective attention capture by auditory alarms”, *J. Human Factors*, vol. 54, no. 5, pp. 826-37, 2012.
[15] E. Sohoglu, J. E. Peelle, R. P. Carlyon, and M. H. Davis, “Top-down influences of written text on perceived clarity of degraded speech”, *J. Exp. Psychol. Hum. Percept. Perform*., vol. 40, no. 1, pp. 186-199, 2014.
[16] A. Zekveld, M. Rudner, I. Johnsrude, J. M. Festen, J. H. M. van Beek and J. Rönnberg, “The influence of semantically related and unrelated text cues on the intelligibility of sentences in noise”, *Ear Hear*., vol. 32, no. 6, pp. e16-E25, 2011.
[17] A. Wingfield, “Cognitive factors in auditory performance: context, speed of processing, and constraints of memory”, *J. Am. Acad. Audiol*., vol. 7, no. 3, pp. 175-182, 1996.
[18] 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.
[19] “Phonetically-balanced 1000 sentences speech database”, NTT Advanced Technology Corporation, 1999.
[20] C. J. Wild, M. H. Davis, and I. S. Johnsrude, “Human auditory cortex is sensitive to the perceived clarity of speech”, *NeuroImage*, vol. 60, pp. 1490 –1502, 2012.
[21] D. Norris, J. M. McQueen, and A. Cutler, “Merging information in speech recognition: Feedback is never necessary”, *Behav. Brain Sci*., vol. 23, pp. 299 –325, 2000.
[22] Q. Summerfield, “Use of visual information for phonetic perception”, *Phonetica*, vol. 36, pp. 314–331, 1979.
[23] S. Kuwano, S. Namba, A. Schick, H. Höge, H.Fastl, T. Filippou and M. Florentine, “Subjective impression of auditory danger signals in different countries”, *Acoust. Sci. Tech.*, vol. 28, no. 5, pp. 360-362, 2007.