Intelligibility of speech spoken in noise/reverberation for older adults in reverberant environments

Nao Hodoshima[1]*, Takayuki Arai* [2]*, and Kiyohiro Kurisu*[3]

[1]Department of Information Media Technology, Tokai University, Tokyo, Japan
[2]Department of Information and Communication Sciences, Sophia University, Tokyo, Japan
[3] TOA Corporation, Hyogo, Japan

hodoshima@tokai-u.jp, arai@sophia.ac.jp, kurisu_kiyohiro@toa.co.jp

Abstract

Speech intelligibility is in general lower for older adults than young adults in reverberant environments such as train stations or airports. We aim at to make speech announcements intelligible in public spaces. Speech spoken in noise, i.e., noise-induced speech, is usually more intelligible than speech spoken in a quiet environment for young people when they are heard in noise, a phenomenon called the Lombard effect. The current study applied this effect for an input of a sound reinforcement system in public spaces. The results of the listening tests conducted by 24 older adults showed that noise/reverberation-induced speech was more intelligible than speech spoken in a quiet environment when they were in reverberant environments (reverberation time of 1.4 s and 2.4 s). The results also showed that the effect of noise/reverberation-induced speech was observed when the recording and listening condition were different. For example, different reverberation times were used between the two conditions and noise-induced speech was intelligible in reverberation. The results suggest that using noise/reverberation-induced speech as an input of a sound reinforcement system might yields higher intelligibility in public spaces.

Index Terms: older adults, speech intelligibility, Lombard effect, reverberation

1. Introduction

The number of older adults is rapidly growing. In Japan, for example, the population of people aged 65 and older was 23.1% of the total population in 2011, which was the highest rate in the world [1]. Older adults have much more difficulty in understanding speech in a noisy and reverberant environment compared with young adults [2]. Therefore, speech announcements used in noisy and reverberant public spaces, e.g., a train station, need to be intelligible enough for older adults.

In the process of speech communication, our speech is modified to make it robust against noise, which is known as the Lombard effect [e.g., 3]. Speech spoken in noise (i.e., noise-induced speech: NIS) are modified in acoustic characteristics compared with speech spoken in a quiet environment, such as increases in intensity, duration, pitch, the first/second formant frequencies (F1 and F2) [e.g., 4, 5]. Also, NIS has higher speech intelligibility than speech spoken in a quiet environment for young people when heard in a noisy environment [e.g., 4, 5]. The Lombard effect has been used in speech enhancement and automatic speech/speaker recognition [e.g., 6-8].

Reverberation is present in most public spaces as well as noise. Both interferers degrade speech intelligibility, whereas their masking patterns are temporally and spectrally different. Noise masks speech simultaneously, while overlap-masking occurs in reverberation, i.e., the energy of preceding phonemes overlaps the following ones [9]. Therefore a correlation exists between a masker and a maskee in the reverberant masking.

In reverberation, we observed the Lombard-like effect in speech production/perception [10, 11]. That is, speech spoken in reverberation (reverberation-induced speech: RIS) increased in intensity, pitch, F1, F2 and decreased in consonant-vowel intensity ratio [10]. As with NIS, RIS is more intelligible than speech spoken in a quiet environment for young people when heard in reverberant environments [11].

As well as the Lombard speech, clear speech was shown to be more intelligible than conversational speech in noise or reverberation [e.g. 12]. Since both the Lombard speech and clear speech improve speech intelligibility under degraded conditions, the difference is that clear speech is spoken under a quiet environment while the Lombard speech is spoken in the presence of noise/reverberation and thus would more reflect human speech adaptation to noise/reverberation. It would be interesting to compare both speech production, but the current paper focuses on the Lombard speech.

The purpose of this study is to make speech announcements intelligible by applying the Lombard-like effect for an input of a sound reinforcement system in public spaces. That is, without further acoustical treatments such as installing extra absorbing materials, our approach is to record/synthesize speech announcements in a way that yield higher intelligibility and radiate them from loudspeakers to public spaces. The current study examined if NIS/RIS is more intelligible for older adults than speech spoken in a quiet environment when that speech is heard in reverberant environments. In addition, this study tested the robustness of NIS/RIS where speaking and listening conditions were different. That is, we tested the intelligibility of NIS in reverberation, which has never been studied before, while NIS was intelligible in noise [4, 5, 10]. We also investigated the intelligibility of RIS when a speaker and a listener hear different reverberation. If NIS/RIS is more intelligible than speech spoken in a quiet environment under such conditions in this study, we do not need to set up exactly the same noisy/reverberant condition for recording/synthesizing speech announcements as for public space where the speech announcements are sent.

A listening test in which older adults listen to NIS/RIS in reverberant environments is described in Section 2. Its results and discussion are described in Section 3, and Section 4 concludes the results.

2. Listening test

2.1. Participants

The participants were 24 native speakers of Japanese (12 males and 12 females; aged between 65 to 78) recruited from the Silver Human Resources Center in Minato ward, Tokyo. Neither of them has worn a hearing aid. Neither of pure-tone thresholds nor speech audiometry of the participants were measured.

2.2. Stimuli

Two native speakers of Japanese [one female (S1) and one male (S2), 20 and 22 years old, respectively] served as speakers. They reported no hearing difficulties and articulation disorders.

Speech materials consisted of 36 target words embedded in a carrier sentence. The target words were four morae (mora is a phonological syllable-like unit in Japanese) and selected from the database of familiarity-controlled Japanese word lists (FW03) [13]. The familiarity of the target words were between 2.5 and 4.0 on a 7-point scale (1 for the most unfamiliar and 7 for the most familiar) [13]. The speech tokens were the same (therefore, the same speakers) as in the listening test on young adults [11].

Speech materials were recorded under three speaking conditions: quiet (Q), noise (N), and reverberation (R) on a computer through a microphone (SHURE, KSM109; condenser, cardioid), an amplifier (PreSonus, DIGIMAX FS), and a digital audio interface (RME, Fireface 800) in a sound-treated room. In the N/R condition, white-noise/reverberant speakers' utterances was presented to the speakers over headphones (SENNHEISER, HDA200; dynamic, closed circumaural type) by Adobe Audition 3.0. The delay caused by the software was less than a few ms. The playback sound did not contain the direct sound and its level was set to -22 dB relative to the speaking level (A-weighted) of the speakers at their ears. Reverberation time (RT) of R condition was 3.6 s at average among octave bands from 125-4000 Hz. The recorded speech materials contained neither noise nor reverberation.

Each speaker read a total of 108 sentences (3 speaking conditions x 36 speech samples). The speakers were instructed to imagine that their speech was being broadcasted to a public space with room acoustics as they heard and to speak as clearly as possible.

In each speaking condition, a carrier sentence was chosen and the target words were embedded in the carrier sentence. This was done to control the effect of the reverberant masking on the target words. The intensity ratio of the carrier sentence relative to each of the target word was normalized within speaking conditions.

The combination of the speaking conditions and reverberant conditions used in the listening test are shown in Table 1. The speech materials were convolved with the two impulse responses (R1 and R2), and RT is 1.4 s and 2.4 s respectively. The impulse responses were selected in order to simulate a public space that has relatively long RT (e.g., subway station, airport). R1 and R2 were made by changing exponential decays of the impulse response used in the recording. The overall intensity of the stimuli was normalized across the speaking conditions and speakers.

Table 1. *Conditions used in the listening test. Speech spoken in quiet (Q), reverberation (R) and noise (N) were presented to participants in two reverberant environments (R1 and R2).*

Speaking condition	Reverberation
Q	R1
	R2
R	R1
	R2
N	R1
	R2

2.3. Procedure

The listening test was carried out in a sound-treated room. The stimuli were presented to the participants diotically over headphones (STAX, SR-303; electrostatic, open circumaural type) through a digital audio interface (TASCAM, US-144MKII) connected to a computer. Two practice trials were held to familiarize the participants with the procedure. The playback level was adjusted to each participant's comfort level. In each trial, a stimulus was presented once, and the participants were instructed to write down what they heard as a target word on their answer sheets. For each participant, 36 stimuli (3 speaking conditions x 2 impulse responses x 2 speakers x 3 set of speech samples) were presented randomly. Combinations of the target words and the listening condition were counter-balanced across the participants.

3. Results and discussion

3.1. Results

The mean percent correct of mora for each speaking condition, each speaker and the average of the speakers at R1 and R2, are shown in Figures 1 and 2 respectively. A 3 x 2 x 2 ANOVA was carried out with speaking condition (Q, R, and N), reverberation (R1 and R2) and speakers (S1 and S2) as repeated variables, and the mora correct rate (from here, correct rate indicates mora correct rate) as the dependent variable by SPSS. We set the significance level at 5%. We made the following hypotheses from the previous studies [4,5,10,11]: (1) Shorter RT have higher correct rate than longer RT, (2) R/N has higher correct rate than Q, and (3) S2 was more intelligible than S1.

The main effect of reverberation was significant (p<0.01), showing that the correct rate significantly decreased as RT increased. The main effect of speaking condition was significant (p<0.01) and a Sidak multiple comparison test showed that significant difference between Q and R (p=0.001) and Q and N (p=0.002). They showed that the correct rates of R and N were significantly higher than that of Q. The main effect of speakers was significant (p<0.01), showing that the correct rate of S2 was higher than that of S1.

In each reverberation, further ANOVA was carried out with speaking condition (Q, R, and N) as repeated variables, and the correct rate as the dependent variable. At R1, no significant main effect was observed. At R2, the main effect was significant (p<0.01) and a Sidak multiple comparison tests showed that significant difference between Q and R (p=0.004) and Q and N

(p=0.001). They showed that the correct rates of R and N were significantly higher than that of Q.

For each speaker, ANOVAs were carried out with speaking condition (Q, R, and N) and reverberation (R1 and R2) as repeated variables, and the correct rate at each reverberation and the average of reverberation as the dependent variable. For S1, the main effect of reverberation was significant (p<0.01), showing that the correct rate significantly decreased as RT increased. The main effect of speaking condition was significant for R2 (p=0.001) and the average of reverberation (p<0.01). A Sidak multiple comparison tests showed that significant difference between Q and N for both conditions (p=0.001). They showed that the correct rates of N were significantly higher than that of Q. At R1, no significant main effect was observed.

For S2, the main effect of reverberation was significant (p=0.002), showing that the correct rate significantly decreased as RT increased. The main effect of speaking condition was significant for R1 (p=0.018), R2 (p=0.01) and the average of reverberation (p<0.01). Results of a Sidak multiple comparison tests showed that significant difference between Q and R for all conditions (R1: p=0.032, R2: p=0.007 and the average: p<0.01). They showed that the correct rate of R was significantly higher than that of Q.

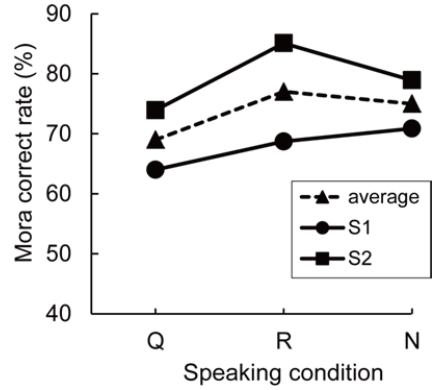

Figure 1: *Mean percent correct of mora for each speaking condition at R1 (RT=1.4 s).*

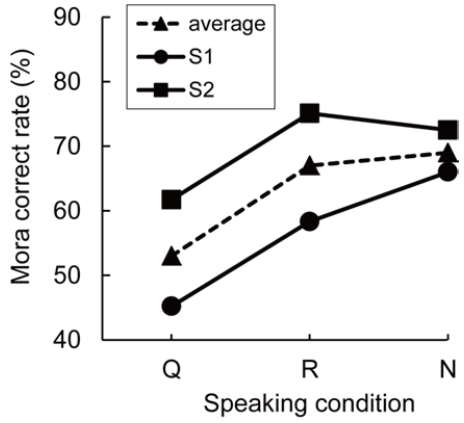

Figure 2: *Mean percent correct of mora for each speaking condition at R2 (RT=2.4 s).*

3.2. Discussion

RIS was more intelligible for older adults than speech spoken in a quiet environment when a speaker was S2 at both reverberant conditions. This might be related to the modification in the acoustic characteristics of RIS over speech spoken in quiet conditions for S2. A t-test showed that pitch of RIS (211Hz) was significantly higher than that of speech spoken in quiet (188Hz) (p<0.01). Those histograms of F0 are shown in Figure 3 and spectrograms and pitch contours of a target "tekigata" are shown in Figure 4 respectively as a reference. F1 (511Hz) and consonant duration (56 ms) were more increased in RIS than those of speech spoken in quiet (489Hz and 35 ms respectively), but the increases were not statistically significant.

When the speaker and listeners heard different reverberation, i.e., difference in RT of the recording and listening conditions were 2.2 s and 1.2 s for R1 and R2 respectively, RIS was more intelligible than speech spoken in Q.

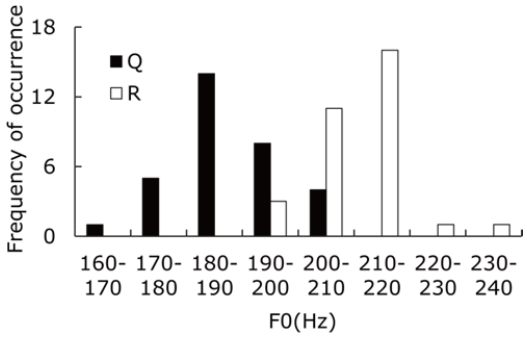

Figure 3: *Histograms of F0 distribution of speech spoken by S2 in quiet (Q) and reverberation (R).*

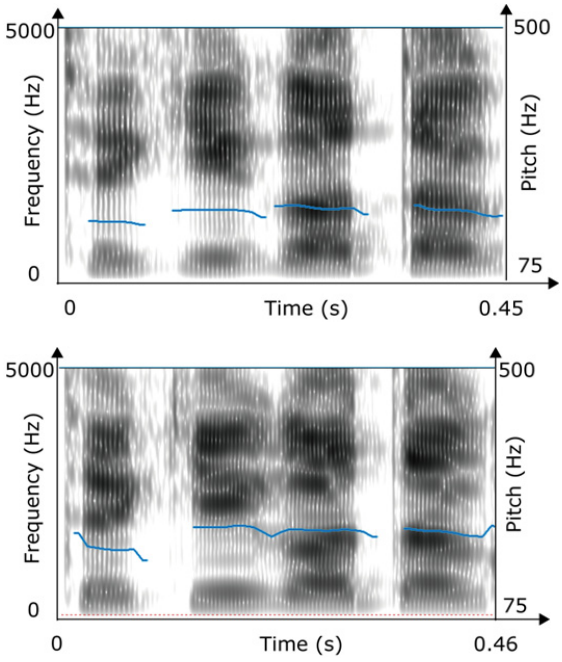

Figure 4: *Spectrograms and pitch contours of a target "tekigata" spoken by S2 in quiet (top) and reverberation (bottom).*

This indicates that a reverberant condition for recording/synthesizing speech announcements is not to be the same as public space where the speech announcements are used.

The effect of RIS is consistent with the previous study on young adults [11] while reverberation time at which RIS was effective was shorter for older adults (1.4 s and 2.4 s) than young adults (2.6 s and 3.6 s) [11]. This agrees with the discussion that older adults are much more affected by temporal smearing of reverberation compared with the effect on young adults with normal hearing [14, 15]. Since young adults and older adults were not tested at exactly the same RTs, we compared the correct rates of young adults at RT of 2.6 s [11] and those of older adults at RT of 2.4 s in this study both for the average of speakers. The correct rates of young adults were 79.1% for speech spoken in quiet and 86.3% for RIS [11]. Both speaking conditions had higher correct rates for young adults than for older adults. However, the improvement in speech intelligibility by RIS over speech spoken in quiet was higher for older adults (13.2%) than young adults (7.2%). It can be said that older adults might have much benefit from RIS compared with young adults.

NIS was more intelligible than speech spoken in a quiet environment when a speaker was S1 at R2. This indicates that the acoustic characteristics enhanced in speech spoken in noise were also intelligible in reverberation under these conditions. Although noise and reverberation have different masking patterns, two interferences share the same acoustic characteristics, such as pitch and formant frequencies, that increase similarly in NIS/RIS compared with speech spoken in quiet conditions [10].

The effect of NIS/RIS varied with speakers. This is the same tendency described in the previous study [11]. This might be because different speakers used different modification styles of speech in noise/reverberation. The effect of NIS/RIS depended on listeners as well. For the older adults, RIS was more intelligible than speech spoken in Q for S2 and NIS was more intelligible than speech spoken in Q for S1. However, RIS was more intelligible than speech spoken in Q for both speakers and NIS was more intelligible than speech spoken in Q for S1 for the young adults. Since there were only two speakers, it is not quite clear that these differences result from speakers, listeners, or the reverberant/noisy conditions. Future investigation should increases the number of speakers and systematically investigates the relation between acoustic characteristics and the intelligibility of NIS/RIS. It would be interesting to find intelligible speakers in noise and reverberation so that we could use such speakers for making speech announcements. On the other hand, if we find the acoustic characteristics of intelligible speakers in noise and reverberation, we could record/synthesize/process speech announcements by enhancing such characteristics for making speech announcements.

4. Conclusions

We found that NIS and RIS were more intelligible for older adults than speech spoken in a quiet environment when the listeners were in reverberant environments (RIS: S2, RT of 1.4 and 2.4 s, NIS: S1, RT of 2.4 s). A possible application of the current findings is to record announcements or develop a speech synthesis system that is intelligible for noisy and reverberant public spaces. Another possible application would be in instructing staff working in public spaces on how to effectively transmit messages to their audiences.

The results indicated that NIS/RIS under the current experimental conditions were robust. That is, RIS was intelligible when the speaker and listeners were in different reverberant situations and NIS was intelligible in reverberation. This implies that the conditions for recording/synthesizing speech announcements do not have to be the same as acoustic conditions of public space.

5. Acknowledgements

This work was partially supported by a Grant-in-Aid for Young Scientists (B) from MEXT (21700203). We are grateful to Hideki Tachibana, Kanako Ueno, and Sakae Yokoyama for providing the impulse response data, to Chikashi Michimata for advice about the experimental design and statistical analysis, and the speakers and listeners who participated in the listening test. The listening test was carried out by Yusuke Miyanaga and Motoki Tsuruo.

6. References

[1] The Cabinet Office “Annual Report on the Aging Society”, Japan, 2011.

[2] Nabelek, A. K. and Robinson, P. K., “Monaural and binaural speech perception in reverberation for listeners of various ages”, J. Acoust. Soc. Am., 71(4):1242-1248, 1982.

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

[4] Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow R. I. and Stokes, M. A., “Effects of noise on speech production: Acoustics and perceptual analysis”, J. Acoust. Soc. Am., 84:917-928, 1988.

[5] Junqua, J.-C., “The Lombard reflex and its role on human listeners and automatic speech recognizers”, J. Acoust. Soc. Am., 93:510-524, 1993.

[6] Hansen, J. H. L., “Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition”, Speech Communication, 20:151-173, 1996.

[7] Hansen, J. H. L. and Varadarajan, V., “Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition”, IEEE Trans. Audio, Speech, Lang. Process., 17:366-378, 2009.

[8] Skowronski, M. D. and Harris, J. G., “Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments”, Speech Communication, 48:549-558, 2006.

[9] Nabelek, A. K., Letowski, T. R. and Tucker, F. M., “Reverberant overlap- and self-masking in consonant identification”, J. Acoust. Soc. Am., 86:1259-1265, 1989.

[10] Hodoshima, N., Arai, T. and Kurisu, K., “Speaker variabilities of speech in noise and reverberation”, IEICE Technical Report, SP2009-69:43-48, 2009. (in Japanese)

[11] Hodoshima, N., Arai, T. and Kurisu, K., “Intelligibility of speech spoken in noise and reverberation”, Proc. ICA, 2010.

[12] Payton, K. L., Uchanski, R. M., and Braida, L. D., “Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing”, J. Acoust. Soc. Am., 95(3): 1581-1592, 1994.

[13] Amano, S., Kondo, T., Sakamoto, S. and Suzuki, Y., “Familiarity-controlled word lists 2003 (FW03)”, The Speech Resources Consortium, National Institute of Informatics in Japan, 2006.

[14] Gordon-Salant, S. and Fitzgibbons, P. J., “Profile of auditory temporal processing in older listeners”, J. Speec Lang. Hear. Res., 42:300–311, 1999.

[15] Fitzgibbons, P. J. and Gordon-Salant, S., “Age effects on discrimination of timing in auditory sequences”, J. Acoust. Soc. Am., 116(2):1126–1134, 2004.