inter.noise
HAMBURG 2016

Effects of speaker's and listener's acoustic environments on speech intelligibility and annoyance

R. Kubo[1], D. Morikawa and M. Akagi
Japan Advanced Institute of Science and Technology, Japan

ABSTRACT
Speech signal modification to increase intelligibility in noisy environments may cause discomfort in quieter environments. Human speech production in noise is known to improve intelligibility in noise. The so-called Lombard effect might cause annoyance in quiet environments. However, the effect disappears in quiet environments. This suggests that humans might modify their voices depending on the acoustic environment considering intelligibility and discomfort. The speech modification algorithms which use rules obtained from human speech production are expected to meet the demand for both of high intelligibility and reducing discomfort in various acoustic environments. This paper examined the influence of the coincidence between speaker's and listener's acoustic environments on speech intelligibility and the level of annoyance and to confirm the expectation. Speech produced by four Japanese speakers in quiet and noisy environments (at non, 66, 78, 90 dB) were presented to Japanese listeners in quiet and noisy environments (at non, 66, 78 90 dB) to measure intelligibility of words and the level of annoyance. The results showed that speech is less intelligible or more annoying when the listener's environment is different from the speaker's one. This supports that mimicking human speech production would meet the demand for high intelligibility and less annoyance.

Keywords: Intelligibility, Annoyance, Speech modification, Lombard effect, Noise

1. INTRODUCTION

Listening to an announcement via a public address system in social spaces (e.g., subway platform, airport terminal) may be affected by noise and/or reverberations. Maintaining intelligibility in such settings is required for speech-based information service systems. One of the simplest solutions to overcome the masking of background noises is to increase in the sound pressure level (SPL) of the announcement to increase signal-to-noise ratio (SNR). However, listening conditions are not stationary. Announcements of large sound volume can be sources of discomfort for people present. Not only noise levels but also spectral characteristics of noise change. In order to maintain intelligibility and acoustical comfort in variable acoustic environments, announcements need to adaptively change to the acoustic environments.

It has been shown that, in noisy environment, speech produced in noise is more intelligible than is speech produced in quiet environments [1, 2]. The speaker in adverse conditions may use a intelligibility-enhancing speaking style. Studies have characterized how the speech modifications (so-called 'Lombard effect') influence different measures of speech production (e.g., intensity, F0, spectral tilt, etc.) compared with "normal" speech. These speech modifications increase the level of annoyance, since the modifications might increase loudness, sharpness, etc., which are related to psychoacoustic discomfort [3]. Humans might adopt "normal" or "Lombard" speaking styles depending respectively on their acoustic environments. Humans would adaptively produce speech sound considering intelligibility and annoyance. If it is true, applying the speech modification algorithms based on human speech production can meet the demand for maintaining intelligibility and reducing annoyance in various acoustic environments.

The primary purpose of this study is to examine whether speakers modify their speech to maintain intelligibility and reduce annoyance according to acoustic environments. For the purpose, this study investigated the influence of the coincidence between speaker's and listener's acoustic environmentso on speech intelligibility and annoyance. The hypothesis is that when the listener's acoustic environment is different from the

[1]email: rkubo@jaist.ac.jp

speaker's one, the speech is less intelligible or its annoyance increases. On the other hand, when speaker's and listener's acoustic environments are coincident, the high level of intelligibility and less annoyance would be expected. The study also examined benefits of the speech modifications other than intensity changes for comparison with SPL increase which is considered as one solution to improve intelligibility in noise.

2. EXPERIMENTS

A factorial design was employed to examine the influences of the coincidence between speaker's and listener's acoustic environments. One factor was the speaker's acoustic environment, and the other was as the listener's acoustic environment. Each factor had 4 levels. Two sets of listening experiments were conducted. In each experiment, two types of tests were conducted. The first test measured intelligibility, and the second one measured the level of annoyance. Experiment 1 was designed to examine any benefits of the entire speech modifications including intensity changes. Experiment 2 was designed to examine the benefits of the speech modifications other than intensity changes. Speech collection was conducted prior to the listening experiments. The target language was Japanese, and native speakers of Japanese participated as speakers in the speech collection. Another native speakers participated in the listening experiments as listeners.

2.1 Speech collection

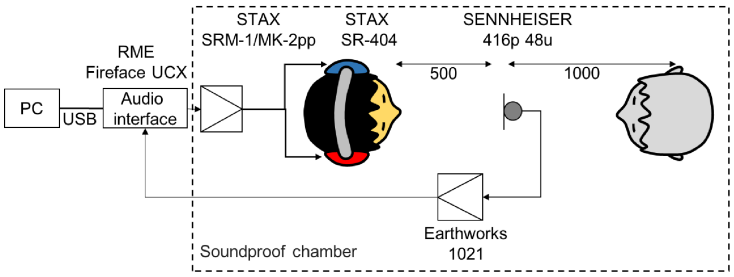
2.1.1 Speech material and noise

Words were selected from familiarity-controlled word lists 2007 (FW07) [4]. Word lists of lowest familiarity rank (20 lists) were used. Each list contained 20 words. Each word consisted of four moras, each mora was V or CV (V: Vowel, C: Consonant). Each word was embedded in a carrier sentence as a target word. Two-channel pink noise of 6000 ms was prepared for noise to be presented to the speakers during the recording.

2.1.2 Speakers

Two males (N10, N11) and 2 females (N12, N13) participated as speakers. None of the participants reported any history of a speech or hearing impairment. All passed a hearing test which was used to test each ear separately at the six frequencies: 250, 500, 1000, 2000, 4000, and 8000 Hz at 20 dB hearing level.

2.1.3 Procedure



ns

Figure 1. Recording setting.

The recording was made in a sound-proof room. Figure 1 provides relative positions of the speaker, microphone, and simulated listener's head (head model). The microphone was located 50 cm from the speaker's mouth, and a head model was located 150 cm away from the speaker's position. Noise was presented to the speaker dichotically using open headphones (STAX SR-404) such that the listeners were able to use auditory feedback from self-produced speech and noise. Four levels of the noise were applied.The noise was not presented to the speaker in the quiet condition (non). The noise was presented at A-weighted SPL of either of 66, 78 or 90 dB at the ears of the speaker in the noisy conditions (66 dB, 78 dB, 90 dB, respectively). Sixteen word lists were prepared per one speaker, and 4 word lists were assigned to each of these acoustic conditions. The recording was conducted in random order blocked by word list. The speaker were instructed to read the

sentences as if talking to a listener at distance of 150 cm from the speaker. The utterances were recorded using a microphone (SENNHEISER 416p 48u), and saved digitally after digitizing at 44.1 kHz with 16-bit quantization resolution. Figure 2 shows the mean SPLs by speaker and recording condition.

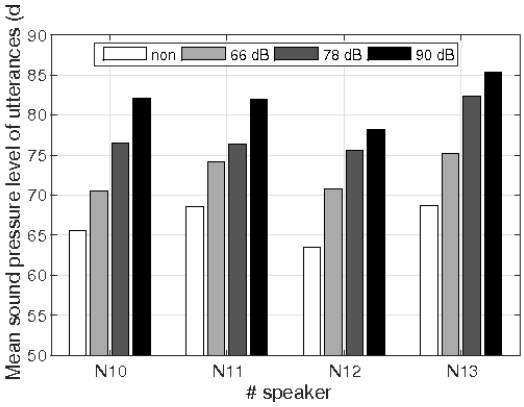


Figure 2. Mean sound pressure level of utterances by speaker, recording condition.

2.2 Experiment 1

2.2.1 Speech intelligibility test

Experiment 1 was designed to investigate any benefits of the entire speech modifications including intensity changes. The stimuli were generated from combinations of the collected speech and the noise signal which was identical to that used during the recording. The amplitudes of speech and noise signals were adjusted in order to simulate the condition that the speaker and listener are in the same or different acoustic environment. The amplitude of the speech signal was adjusted so that each speech was presented to the listener at the same SPL as the one at the ears of the head model during the recording. The amplitude of the noise signal was adjusted in three levels so that the noise was presented at A-weighted 66, 78 or 90 dB at the ears of the listener. Then, the speech and noise signals were combined so that the noise started 300-600 ms earlier than the speech, and lasted for 6000 ms. Speech signals without noise were prepared for a quiet condition. One of the word lists (20 words) was assigned to an intersection between speaker's and listener's conditions (16 intersections), and then a total of 320 stimuli were prepared per listener. Three sets of stimuli were prepared, and one of the sets was chosen for each listener.

Thirteen graduate students participated. Hearing test was administrated as in the speech collection. Participants were tested in a sound-proof room with a background noise of less than A-weighted 21 dB sound pressure level. All tests were administered on a personal computer to present stimuli and to collect responses. Stimuli were presented from the personal computer through an audio interface (RME Fireface UCX) and headphones (STAX SR-404). The stimuli were presented once in random order. The participants heard a stimulus and were required to type what they heard in katakana characters (Japanese phonetic symbols).

2.2.2 Annoyance rating

Six stimuli per each intersection between speaker's and listener's conditions were selected from stimuli which were used in the intelligibility test. The stimuli for intersections where the noise level for speakers was lower than the one for listeners (i.e., the speech was recorded in noise at 66 dB, and combined with noise at 90 dB) were excluded in this experiment because there was a possibility that the speech was inaudible.

Eight graduate students participated. Hearing test was administrated as in the intelligibility test. Participants were tested in the same environment as that in the intelligibility test. The stimuli were presented in random order. The participants heard a stimulus and were required to rate the degree of annoyance of the utterance using a five point scale (1: not annoying – 5: annoying).

2.2.3 Results & Discussion

The answers from the intelligibility test were analyzed for mora intelligibility scores by calculating the percentage of identified mora. A repeated measures ANOVA was conducted with the arcsine-transformed scores

as the dependent variable, with speaker condition (non vs. 66 dB vs. 78 dB vs. 90 dB) and listener condition (non vs. 66 dB vs. 78 dB vs. 90 dB) as the within-subject factors. The analysis revealed a significant interaction ($F(9, 108) = 32.35, p < .01$) between speaker and listener conditions. A pairwise comparison with Bonferroni correction showed no difference among speaker conditions when the listener condition was quiet (non). On the other hand, speech recorded in high levels of noise (78, 90 dB) was more intelligible than that recorded in a quiet condition (non) when the listener condition was noisy.

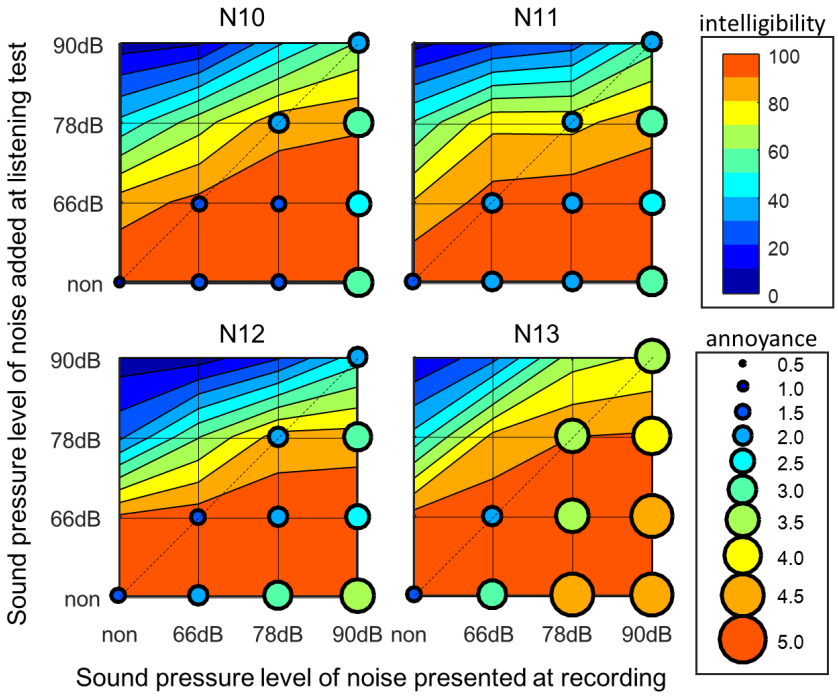


Figure 3: Estimated intelligibility (contour plot), and annoyance rating (bubble chart) by speaker.

Figure 3 contains contour plots of estimated intelligibility by speaker. The mean scores from the annoyance rating are displayed as bubble charts over the intelligibility contour to investigate the influences of the speaker's and listener's acoustic environments.

The contour lines for intelligibility rise to the right. The finding is consistent with previous studies which have reported that speech produced in noise is more intelligible than speech produced in quiet. In addition, the estimated intelligibilities on the main diagonal less than 78 dB are approximately over 80 %. It suggests that the speech is fairly intelligible when the listener's and speaker's acoustic environments are the same. The improvement of intelligibility may at least partly depend on the increased intensity of the speech. As shown in Figure 2, the SPL of the speech increased as the noise level increased during the recording. The speakers could modify their speech to increase SNR using auditory feedback from self-produced speech and noise in the surrounding environment. However, the SNR is not high enough when the noise is at 90 dB. It might be difficult to produce more than 90 dB due to the production mechanism. Past the certain point, shouting which decreases phonetic information might have occurred [5]. However, it should be noted that the speech recorded in the same condition (90 dB) is more intelligible than speech recorded in other conditions. The intelligibility can be high when the listener's condition is quieter than or equal to the speaker's one.

Second, the degree of annoyance increases from left to right along the vertical axis, indicating that annoyance increases as the speaker's condition becomes noisy. Furthermore, it can be seen that annoyance decreases from bottom to top along the horizontal axis, indicating that annoyance decreases as the listener's condition is closer to the speaker's one. Both vertically and horizontally, the least annoyance is mostly located on the main diagonal (where the speaker's and listener's conditions are the same). Annoyance can be decreased when the listener's acoustic environment is equal to the speaker's one.

As expected, taken together, the relatively high level of intelligibility and less annoyance were obtained when the speaker's and listener's acoustic environments are coincident.

2.3 Experiment 2

2.3.1 Speech intelligibility test

In Experiment 2, in order to investigate the benefits of the speech modifications other than intensity changes, the speech signals were normalized. All stimuli prepared for a listening condition were normalized so that the targets words had equal rms energy over the stimuli regardless of the recording condition. For each level of listening condition (noise level added at the listening test), all speech signals of each speaker were normalized by using the mean value of SPL of speech produced in the same noise level (Figure 2) as the standard, then combined with noise. Then, SNR conditions were postulated to be equal across all stimuli in a listening condition, and speaker. As in Experiment 1, the amplitude of the noise signal was adjusted in three levels so that the noise was presented at A-weighted 66, 78 or 90 dB at the ears of the listener. Speech signals without noise were prepared for a quiet condition, then a total of 320 stimuli were prepared per listener.

Eight graduate students participated. Hearing test was administrated as in Experiment 1. The procedures were identical to those of the intelligibility test of Experiment 1 except for the stimuli.

2.3.2 Annoyance rating

Stimuli which were prepared for the intelligibility test were used. Six stimuli for each condition were selected. Because of the rms-normalization, although the stimuli which the noise level during recording was higher than the one during listening test were excluded in Experiment 1, the exclusion was not done in Experiment 2.

Five graduate students participated. Hearing test was administrated as in Experiment 1. The procedures were identical to those of the annoyance rating of Experiment 1 except for the stimuli and number of trials.

2.3.3 Results & Discussion

The answers from the intelligibility test were analyzed for mora intelligibility scores by calculating the percentage of identified mora. A repeated measures ANOVA was conducted with the arcsine-transformed scores as the dependent variable, with speaker condition (non vs. 66 dB vs. 78 dB vs. 90 dB) and listener condition (non vs. 66 dB vs. 78 dB vs. 90 dB) as the within-subject factors, but no significant effect was found.

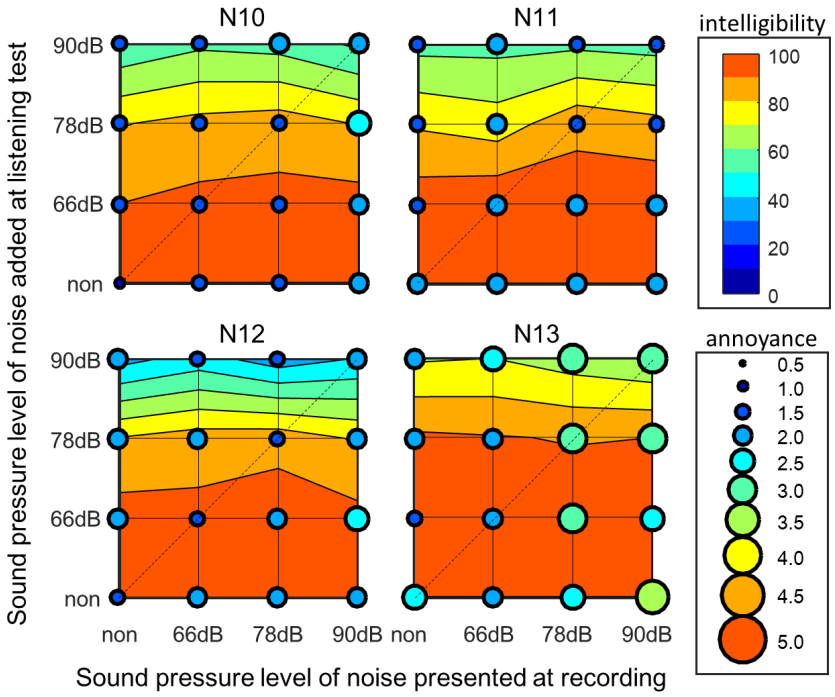


Figure 4: Estimated intelligibility (contour plot), and annoyance rating (bubble chart) by speaker for normalized speech.

Figure 4 displays each speaker's estimated speech intelligibility as contour plots, and annoyance level as bubble charts to investigate the influences of the speaker's and listener's acoustic environments. The contour

lines for intelligibility of 80 and 90 % form small rising-falling peaks around 78 dB except one speaker (N13). It suggests that speech produced in noise can be more intelligible than speech produced in quiet. Although the increase is not significant, it suggests that listeners benefited from speech modifications other than intensity changes. Second, as for annoyance, it is possible to see the pattern which shown in Experiment 1, even though it does not cover all speakers. In both vertical and horizontal axes, the annoyance on the main diagonal line tends to be smaller than or equal to others. This trend is depended on the speaker, speaker N12 shows a relatively clear pattern of this, while speaker N13 does not.

Taken together, for three of four speakers, the relatively high level of intelligibility and less annoyance can be obtained when the speaker's and listener's acoustic environments are closer, even after normalization. It should be noted that the speech normalization was done simply by rms-normalization over the entire word part. Some aspects of speech modifications (i.e, spectral sharpening) are averaged out in the normalization. Future work should investigate the influence of these speech modifications on intelligibility and annoyance.

Interestingly, after normalization, speech produced in a quiet condition can be more annoying than is one produced in a coincident condition with listener's one. As can be seen in N12, the amplified "normal" speech is not intelligible but annoying. It suggests potential risks in the use of amplified normal speech, which might contribute little to intelligibility but increase the level of annoyance.

3. GENERAL DISCUSSION & CONCLUSION

The results from Experiment 1 implies that when speaker's and listener's acoustic environments are coincident, higher intelligibility and less annoyance can be obtained. The results from Experiment 1 and 2 not only imply the contribution of intensity to intelligibility in noise but also suggest the benefits of speech modifications other than intensity changes. They also suggest the risks of amplifying normal speech, which contributes little to intelligibility and increase the level of annoyance. Extracting effective speech modifications in noisy environments, and emphasizing the modifications would have potential benefits to effective public address in various noisy environments. It remains for future research to investigate which acoustic feature contributes to the improvement in intelligibility and/or annoyance.

These finding support that speakers modify their speech to maintain intelligibility and reduce annoyance according to acoustic environments. Based on the assumption that humans adaptively produce speech sounds considering intelligibility and annoyance according to the environment, applying speech modifications in human production to announcements is expected to meet the demand for maintaining intelligibility and reducing annoyance in various acoustic environments.

ACKNOWLEDGMENTS

A part of this research was supported by SECOM Science and Technology Foundation.

REFERENCES

[1] W. Van Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes. Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America*, 84(3):917–928, 1988.

[2] H. Lane and B. Tranel. The lombard sign and the role of hearing in speech. *Journal of Speech, Language, and Hearing Research*, 14(4):677–709, 1971.

[3] H. Fastle and E. Zwicker. *Psychoacoustics: Facts and Models*. Springer, 2006.

[4] K. Kondo, S. Amano, Y. Suzuki, and S. Sakamoto. Japanese speech dataset for familiarity-controlled spoken-word intelligibility test (FW07). NII Speech Resources Consortium, 2007.

[5] D. Rostolland. Intelligibility of shouted voice. *Acta Acustica united with Acustica*, 57(3):103–121, 1985.