Study on improvement of speech intelligibility under reverberant environment focusing on modulation spectrum

1810183 Shota Morita

Reverberation various reverberation times exists in indoor public spaces such as stations and concert halls. In the presence of reverberation such as evacuation guidance voices during disasters, it is necessary to accurately transmit voices to people in the facility.

When speaking in noisy environments, humans change voice in response to noise to improve speech intelligibility. This phenomenon is known as the Lombard effect. Similar compensation movements may be performed in reverberant environments. To date, various studies have been conducted on speech in reverberant environments. However, it has not been clarified yet features that are important for accurately transmitting speech in the deformation of speech in a reverberant environment.

Arai et al. that speech uttered in a reverberant environment tended to have higher intelligibility in a reverberant environment than speech uttered in a silent environment. In addition, they report that consonants are coordinated by steady-state suppression corresponding to a vowel for speech, and the intelligibility under reverberation is significantly increased.

Kubo et al. speech uttered under a total of four conditions of reverberation environment of quiet sound and reverberation time (1 second, 3 seconds, 5 seconds), and conducted a listening experiment with reverberation convolved. It was reported that some speakers tended to have higher speech intelligibility in reverberant environments for longer. Also, focusing on the formant frequency of the uttered voice of the same speaker, analysis is performed, and the vowel space is expanded when the reverberation time during utterance becomes longer, and the first formant in the onset (0 to 25 %) of the vowel section The slope of the transition between (F1) and the second formant (F2) is the steepest in the non-giving condition, but the transition of the formant is steeper when the reverberation time is 5 seconds compared to the reverberation time of 1 second when speaking. Has been reported.

However, in each case, only the results on the frequency axis are taken into account, and the time variation of the utterance is not considered. In this paper, temporal envelope of speech in order to clarify acoustic features that are important for improving speech intelligibility in reverberant environments. By analyzing the modulation spectrum of speech in various reverberation environments, focusing on F1 and F2, which are considered important in previous research,This paper was found features that are considered to be

related to the improvement of speech intelligibility. A listening experiment is performed to determine whether the obtained features have contributed to the improvement of speech intelligibility in a reverberant environment.

Speech uttered in a space resembling a reverberation environment with the same reverberation time (T60) as Schroeder's RIR to investigate whether humans are deforming speech to compensate for information lost by reverberation Speech (40 voices per condition), whose intelligibility tended to increase in the reverberant environment corresponding to the extension of the reverberation time (1 s, 3 s, 5 s), was used for analysis. The frequency bands corresponding to vowels F1 and F2 were extracted with a bandpass filter, and the modulation spectrum was derived to obtain the average value of 40 samples. The results showed that the modulation spectrum of a speech with a long reverberation time tends to be larger than that of a speech with a short reverberation time. In addition, the modulation spectrum around the modulation frequency of about 10 Hz is particularly large, and it is considered that there is a possibility that compensation motion is performed for the modulation frequency component that is damaged when the reverberation time is long.

Next, in order to evaluate the relationship between the utterance deformation and the modulation spectrum in the reverberant environment obtained by the analysis, the intelligibility survey was conducted by listening experiments. The stimulus sound creation method used in the listening experiment was resampled to a sampling frequency of 16000 Hz, and then divided into 64 channels (BandWidth: 125 Hz) using an acoustic filter bank. After obtaining the power envelope from the output of each channel and filtering it to 0 and 1 using Voice Activity Detection (VAD), it is further divided into 64 channels (BandWidth: 1.95Hz) by the Modulation filterbank and specific modulation by the gain control The frequency component is raised and returned to voice. This time, five native speakers of Japanese speak the three most important words in the ATR database and use the voices of the three-mora words. The frequency bands corresponding to vowels F1 and F2 (male: 300 to 2000 Hz) A sound (4 dB up) in which the modulation frequency of the power envelope of the filter envelope: 2 to 16 ch) was around 4 and 10 Hz (Modulation filter bank: 2, 3 ch, 5 and 6 ch) was raised by 4 dB (8 dB up) and a sound was raised by 8 dB. For the experimental stimulus, three types of reverberation (1 s, 3 s, 5 s) convoluted with the original voice, 4 dB up, and 8 dB up. An experiment was performed using this speech to determine the intelligibility of words in mora units. Nine native speakers of Japanese in their 20 s participated in the listening experiment. The stimulus sound was blocked at each reverberation time, and the sound in the block was presented randomly.

Although there was no difference in intelligibility between the reverber-

ation time of 5 s and 3 s, the results showed that the intelligibility of 4 dB up was higher than that of the original sound under the condition of the reverberation time of 3 s. From this fact, it was suggested that the utterance deformation of raising the modulation frequency component for the reverberant environment might be significant, and would lead to the improvement of intelligibility.

This study shows that humans can improve intelligibility by increasing the modulated frequency content of speech in reverberant environments, and perform similar processing to increase speech intelligibility in reverberant environments.