Improvement of speech intelligibility by reallocation of spectral energy

Reiko Takou, Nobumasa Seiyama, Atsushi Imai

NHK (Japan Broadcasting Corporation)
Science and Technology Research Laboratories, Tokyo, Japan
{takou.r-go, seiyama.n-ek, imai.a-dy}@nhk.or.jp

Abstract

This report describes a method of speech modification to improve the intelligibility of speech. The spectral energy is reallocated by maintaining and emphasizing acoustical features important for speech perception. The method consists of three components: flattening of the spectral tilt, enhancement of the spectral contrast, and holding of the low-frequency region. The results of this speech intelligibility test show that this method improves the speech intelligibility for all conditions of noise and SNR employed in the test.

Index Terms: speech intelligibility, speech enhancement

1. Introduction

The improvement of speech intelligibility is greatly needed for various devices and situations such as a noisy environment so that speech may be heard correctly and information can be obtained from speech. This is important for not only hearing-impaired or elderly people but also people with normal hearing in various situations.

In this study, we develop a speech modification method that improves the intelligibility of speech. The method reallocates the energy of a spectrum while maintaining the spectral features that are expected to be important cues for speech perception. Spectral contrast, spectral tilt, and low-frequency harmonics are controlled in the method.

Enhancement of the spectral contrast involves emphasizing the peaks and attenuating the valleys in a spectrum to increase the difference between the levels of the peaks and valleys. A spectral contrast enhancement method that suppresses the noise in spectral valleys and emphasizes spectral peaks has been proposed to improve speech perception for hearing-impaired people [1]. The method, which is intended to reproduce the two-tone suppression phenomenon for cochlear-implant processors, is able to enhance spectral contrast [2]. Neither approach requires the correct extraction of formant frequencies and harmonics, and both methods are considered to be tolerant of distortion and useful for enhancing speech intelligibility.

Spectral tilt is flattened to accurately extract formant frequencies in many cases. In terms of intelligibility, Lu and Cooke have demonstrated that flattening of the spectral tilt, which is one of the characteristics observed in Lombard speech, contributes greatly to the increased intelligibility of speech in noise [3]. On the basis of the above discussion, we employ a process to flatten the spectral tilt in our method. We also aim to enhance the energy of higher frequencies of consonants to improve the perception of such consonants by flattening them.

Low-frequency harmonics include the important cue of the fundamental frequency (F0). Regarding the contribution of F0 to intelligibility, Binns and Culling have indicated that the F0 contour shape is important in speech perception and that the effect is greater in single-talker interference than in speech-shaped noise [4]. The contribution of the F0 contour is expected to change with the type of interference and noise. However, we employ the process of holding the low-frequency region including low-frequency harmonics to help listeners perceive and use the F0 cue and F0 contour.

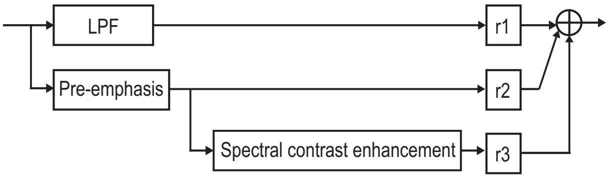

Figure 1: *Block diagram of the proposed method.*

This report describes a method of speech modification consisting of the above three processes and the results of evaluating the efficiency of the method by an intelligibility test that was held by the Hurricane Challenge [5].

2. Method

Figure 1 shows a block diagram of our speech modification method. The original speech is filtered by a low-pass filter (LPF) or is processed by pre-emphasis. The pre-emphasized speech is used without further enhancement or is processed by spectral contrast enhancement and power compensation. The speech enhanced by spectral contrast enhancement is band-limited in the enhancement process. Next, the gains of these three processed speeches are controlled and summed to produce an output speech. The speech data has a 16 kHz sampling rate and 16-bit resolution, and is provided by the Hurricane Challenge. From the spectral contrast enhancement carried out before the final summation, the speech data is processed on a frame-by-frame basis and then speech frames are overlap-added. The speech data for each frame is 32 ms.

This method reallocates the spectral energy balance and the processing is performed with fixed parameters. We did not modify the speech in the time domain, for example, by time stretching or compression, and we did not change the processing or the parameters between voiced and unvoiced speech. In the following sections, we describe the components of flattening of the spectral tilt, enhancement of the spectral contrast, holding of the low-frequency region, and other processes involved in the proposed method.

2.1. Flattening of spectral tilt

The spectral tilt is flattened by pre-emphasis with the first-order Finite Impulse Response (FIR) filter whose coefficient is equal to -0.97. The entire bandwidth of the spectrum is pre-emphasized. The average power of a sentence speech with pre-emphasized is adjusted to the same power as before the pre-emphasis.

2.2. Enhancement of spectral contrast

We applied the process proposed by Turicchia and Sarpeshkar [2] to enhance the spectral contrast. The reason was that this process keeps the relation among the levels of the formant peaks stable and thus, the sound quality is also stable. The process employs the filter bank and the companding architecture. The processing of a single channel consists of pre-filter, compression, post-filter, and expansion.

In our preliminary examination, we selected the parameters $q1 = 2$ for the pre-filter, $q2 = 12$ for the post-filter, $n1 = 0.3$ for the compression, $n2 = 1$ for the expansion, and $w = 40$ the parameter of the time constant of the envelope detector (ED) [6], with the other parameters being the same as those in [2]. Thus, the spectrum of the speech data in the frequency range of 250-4000 Hz was processed, and then the processed speech was band-limited.

On the basis of our preliminary examination, we expected that a processing region with a frequency of above 4000 Hz would make a small contribution to increasing the energy efficiency and improving the intelligibility. We also expected that processing a region in which formant peaks exist would contribute to vowel perception. Power compensation was also required to adjusts the average power of the speech frame with enhanced spectral contrast to the same power as before the enhancement.

2.3. Holding of low-frequency region

Because we considered that the low-frequency region including harmonics is important for intelligibility, the spectrum in the low-frequency region was held by a LPF. Actually, the power of the region below 400 Hz tends to be reduced by the process used to enhance the spectral contrast in section 2.2. For this reason, speech data is filtered by a LPF with a 400 Hz cutoff and a 60th-order FIR filter using a Blackman window.

2.4. Balance of energy

Finally, the gains of the processed speech after LPF, pre-emphasis, and spectral contrast enhancement are controlled to the appropriate level and then summed to obtain the output speech. After a preliminary examination, we set $r1 = r2 = r3 = 1$ as the parameters. The reason for summing the gains of the processed speech is to fill the spectrum above 4000 Hz and compensate for the sound quality of the processed speech, which is deteriorated upon enhancing the speech contrast.

3. Evaluation results

The intelligibility of the speech modified by our method was evaluated in accordance with the Hurricane Challenge. The test conditions were two maskers (competing speech, speech-shaped noise), three SNRs (high, middle, low). Figure 2 shows the result of modified speech (Proposed) and unmodified speech (Plain) under all conditions. The results indicate that the intelligibility score is improved under all conditions.

The increase in the intelligibility score, i.e., the difference between the intelligibility score for Proposed and that for Plain shows that a low SNR has a greater improvement in the case of SSN, whereas a middle SNR has a greater improvement in the case of CS. Moreover, the relative gain for SSN is larger than that for CS. Before giving a full discussion of these results, a detailed analysis of the evaluation data is required in future works.

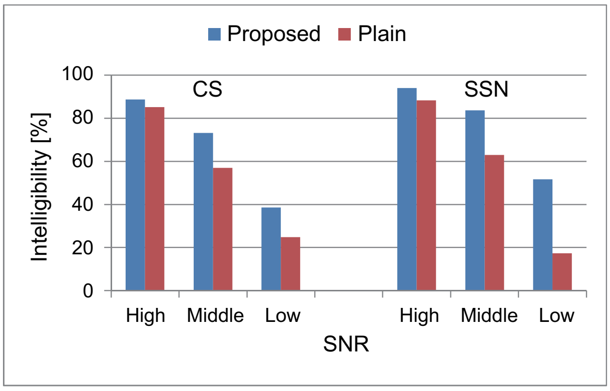

Figure 2: *Intelligibility score.*

4. Discussion

We confirmed the effectiveness of the proposed method in improving speech intelligibility from the intelligibility test conducted in accordance with the Hurricane Challenge.

The results show that the intelligibility gain for SSN is higher than that for CS. Binns and Culling have indicated that the effect of the F0 contour shape on speech perception is greater in the case of single-talker interference than for speech-shaped noise [4]. On this basis, we expected that holding the low-frequency region would improve the tolerance of speech intelligibility to the CS condition. However, it is unclear from this result whether such an effect actually occurred. We need to carry out a detailed analysis of the results before giving a further discussion. The proposed method controls the energy of the spectrum in the frequency domain rather than controlling the speech in the time domain. It is considered that such control is more effective in cases of stationary interference such as SSN than for nonstationary CS. Lu and Cooke have reported that spectral modification alone cannot account for the entire increase in intelligibility of Lombard speech and that modifications in the time domain may also contribute to intelligibility [3]. Thus, we also expect that some control in the time domain will improve the intelligibility.

5. Conclusions

We have proposed a method of speech modification for improving the intelligibility of speech in noise. From the results of an intelligibility test in accordance with the Hurricane Challenge, the intelligibility was found to be improved under all six conditions in the test, allowing us to confirm the effectiveness of our method for increasing intelligibility. The results also show that a greater improvement occurred at a lower SNR and that the improvement was greater under the SSN condition than for CS. We will analyze these results in more detail in future work. We will also examine the effectiveness of our method in various situations. Moreover, a reexamination of the processing cost is required to enable the method to be used with various speech data, speakers, and listening situations.

6. References

[1] Baer, T., Moore, B. C. J. and Gatehouse, S., "Enhancement of spectral contrast of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times", J. Rehabil. Res. Dev., 30(1): 49–72, 1993.

[2] Turicchia, L. and Sarpeshkar, R., "A bio-inspired companding strategy for spectral enhancement", IEEE Trans. Speech Audio Proc., 13(2): 243–253, 2005.

[3] Lu, Y. and Cooke, M., "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise", Speech Commun., 51(12): 1253–1262, 2009.

[4] Binns, C. and Culling, J. F., "The role of fundamental frequency contours in the perception of speech against interfering speech", J. Acoust. Soc. Am., 122(3): 1765–1776, 2007.

[5] http://listening-talker.org/hurricane/

[6] Bhattacharya, A. and Zeng, F. G., "Companding to improve cochlear-implant speech recognition in speech-shaped noise", J. Acoust. Soc. Am., 122(2): 1079–1089, 2007.