Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 78 (2016) 11–18

www.elsevier.com/locate/specom

Effect of processing-based and microphone-based noise reduction algorithms on intelligibility-related acoustic features: A parametric investigation study

Heepyung Kim[a,1], Kyoung Won Nam[a,1], Jinryoul Kim[b], Sunhyun Yook[c], Dong Pyo Jang[a], In Young Kim[a,*]

[a] Department of Biomedical Engineering, Hanyang University, Seoul 133-791, Republic of Korea
[b] Department of Otolaryngology-Head and Neck Surgery, Samsung Medical Center, Seoul 135-710, Republic of Korea
[c] Department of Medical Device Management & Research, Sungkyunkwan University, Seoul 135-710, Republic of Korea

Received 16 July 2015; received in revised form 27 November 2015; accepted 4 January 2016
Available online 12 January 2016

Abstract

It is known that processing-based noise-reduction (PNR) algorithms cannot significantly improve speech intelligibility in noisy situations; however, there have been a few studies that have attempted to explain why. In this study, we performed a word-based parametric investigation to determine the acoustic features that are essential for speech intelligibility that are deteriorated by environmental noises and cannot be sufficiently restored by PNR-processing. Thirty-six Korean bi-syllabic words were utilized for four noise types – babble, car, white, and traffic – and three noise intensities with −5, 0, and +5 dB signal-to-noise ratios. Experimental results demonstrated that among the six word-based acoustic features, two features – amplitude modulation (AM) in the range of 4–16 Hz and spectral balance (SB) – commonly showed relatively high correlations (>0.60) with and high contribution ratios (>30%) to the measured and estimated intelligibility after performing one to one analysis and multi variate analysis respectively; however, the AM and SB values were not significantly restored after PNR-processing but were significantly restored after applying a comparative microphone-based noise-reduction (MNR) algorithm. In spite of several limitations that need to be addressed in future studies, we expect that improving conventional PNR algorithms to reinforce the performance of AM and SB restoration may enhance speech intelligibility in noisy situations.

Keywords: Acoustic feature; Prosodic cue; Spectral cue; Speech enhancement; Speech intelligibility.

1. Introduction

The improvement of the perceptual aspect of a target's speech signal by attenuating or even eliminating undesired background signals is called speech enhancement (Loizou, 2013). Such speech enhancement techniques aim to reduce or eliminate environmental acoustic interferences, such as ambient noises and reverberation, in order to improve the

Abbreviations: PNR, processing-based noise-reduction; MNR, microphone-based noise-reduction; AM, amplitude modulation in the range of 4–16 Hz; SB, spectral balance; FRT12, formant frequency ratio F1/F2; FRT23, formant frequency ratio F2/F3; F0, fundamental frequency; EO1, energy over 1 kHz; DHA, digital hearing aid; NH, normal-hearing; HI, hearing-impaired; ISNR, input signal-to-noise ratio; KLT, generalized subspace algorithm with embedded pre-whitening; pKLT, perceptual KLT; logMMSE, log minimum mean square error; logMMSE-SPU, logMMSE with speech presence uncertainty; specsub, spectral subtraction; SS-RDC, spectral subtraction based on reduced delay convolution; MB, multiband spectral subtraction; Wiener-WT, Wiener filtering based on wavelet-threshold multi-taper spectra; Wiener-AS, Wiener filtering based on a priori SNR estimation; DM1, first-order directional microphone; SRT, speech recognition threshold; CVC, consonant-vowel-consonant; PCA, principal component analysis.

* Corresponding author. Tel.: +82 2 2291 1713, +82-2-2220-0698; fax: +82 2 2220 4949.

E-mail addresses: heepyung@bme.hanyang.ac.kr (H. Kim), kwnam@bme.hanyang.ac.kr (K.W. Nam), jinryoul@gmail.com (J. Kim), shyook@skku.edu (S. Yook), dongpjang@gmail.com (D.P. Jang), iykim@hanyang.ac.kr (I.Y. Kim).

[1] Heepyung Kim and Kyoung Won Nam contributed equally to this paper and should therefore be regarded as equivalent first authors.

http://dx.doi.org/10.1016/j.specom.2016.01.001
0167-6393/

quality and intelligibility of a speech signal, and are often used in various voice communication fields such as mobile phones, voice over Internet protocol services, teleconference systems, and digital hearing aids (DHAs) (Benesty et al., 2005; Loizou, 2013). For several decades, various processing-based (i.e. using single microphone input) noise-reduction (PNR) algorithms – e.g. the subspace method, minimum mean square error (MMSE), logMMSE, and Wiener filtering – and microphone-based (i.e. using multiple microphone inputs) noise-reduction (MNR) algorithms – e.g., automatic, adaptive, and multi-channel directionality approaches – have been developed and applied to enhance the performance of various communication systems in noisy environments (Lim, 1978; Ephraim and Malah, 1984, 1985; Scalart, 1996; Hu and Loizou, 2003). At the same time, many trials in laboratory and real-world environments have been attempted to evaluate the actual clinical efficacy of such PNR and MNR algorithms in terms of speech quality and both objective and subjective speech intelligibility (Walden et al., 1999; Bentler et al., 2008). In these previous trials, the clinical efficacy of the MNR algorithms in the presence of environmental noises was identified in terms of both speech quality and intelligibility for both normal-hearing (NH) and hearing-impaired (HI) people. In addition, the clinical efficacy of the PNR algorithms in noisy environments was verified in terms of speech quality for both NH and HI people. When Hu and Loizou (2007b) evaluated the clinical effects of 13 popular PNR algorithms based on a subspace method, spectral subtraction, Wiener filtering, and a statistical model using 32 NH volunteers in a laboratory environment, the speech quality showed statistically significant improvements in situations involving car, street, and train noises. However, whether these PNR algorithms can also improve speech intelligibility in NH or HI people in the presence of environmental noises has been a topic of consistent debate, especially in HA fields. For example, when Lim (1978) evaluated the clinical effect of a correlation subtraction-based PNR algorithm in noisy situations (speech-in-white noise) with −5, 0, and +5 dB input signal-to-noise ratios (ISNR) with seven NH subjects in a laboratory environment, there was no significant improvement in speech intelligibility in any of the ISNR conditions. In addition, when Hu and Loizou (2007a) performed the speech recognition threshold test using eight PNR algorithms and 40 NH subjects, a small improvement in speech intelligibility (<10%) was observed under stationary noise; however, no significant improvement in speech intelligibility was observed under non-stationary noise. Furthermore, when Bentler et al. (2008) turned on and off the intrinsic PNR function of a commercial behind-the-ear type HA (AxentTM; Starkey, Minnesota, USA) and evaluated the speech intelligibility, ease of listening, and listening comfort using 32 HI subjects in both laboratory and real-world environments, there were significant improvements in the ease of listening and listening comfort when the PNR function was turned on, but there was no significant improvement in speech intelligibility. Based on these previous studies, Loizou and Kim (2011) tried to identify the reason why PNR algorithms cannot seriously improve speech intelligibility in the presence of environmental noises. In their article, they showed that the speech distortion caused by the inaccuracies in voice activity detection and noise-estimation algorithms induces the deterioration of speech intelligibility in noisy situations by affecting the acoustics and phonetics; however, they could not determine a reason for this speech intelligibility effect.

In this study, we investigated the effect of PNR and MNR algorithms on the acoustic features that are known to be related to speech intelligibility, and tried to determine the specific features that are involved in the performance difference between PNR and MNR algorithms in noisy situations.

2. Materials and methods

2.1. Selection of acoustic features

Human auditory organs perceive sounds by analyzing the morphological (temporal or spectral) characteristics of the acoustic signal input (Zwitserlood et al., 2000); therefore, discriminating delicate variations in the morphological characteristics of the signal – e.g. /tala/ and /taja/ – are important for speech intelligibility. In this study, we assumed that conventional PNR algorithms could not sufficiently compensate for or improve one or more acoustic features that are highly correlated with speech intelligibility during NR-processing, and as a result, could not significantly improve speech intelligibility. Although some debates still remain (Amano-Kusumoto, 2010), based on this assumption, we selected several acoustic features known to be correlated with speech intelligibility. For example, Brown and Bacon (2010) and Binns and Culling (2007) demonstrated the relationship between fundamental frequency (F0) and speech intelligibility. Krause and Braida (2004) and Hazan and Markham (2004) demonstrated the relationship between an energy increase in the long-term average spectrum in the range of 1–3 kHz (EO1) and speech intelligibility. Drullman et al. (1994a, 1994b) demonstrated the relationship between the amplitude modulation in the frequency range of 4–16 Hz (AM) and speech intelligibility. Amano-Kusumoto (2010) demonstrated the relationship between the spectral balance (SB), formant frequency ratios of F1/F2 (FRT12), formant frequency ratios of F2/F3 (FRT23) and speech intelligibility. By referring to the articles above, we selected six acoustic features for words – F0, EO1, AM, SB, FRT12, and FRT23 – in this study. The values of the six selected features were calculated by referring to the methodologies and equations in the articles mentioned above. In summary, first, long-term average speech spectrums (LTASSs) of original and noisy signals were calculated (2048 points FFT, 50% overlap, and Hamming window). Then, (1) F0 was determined by entering the LTASS to a cepstrum F0 estimation algorithm (Amano-Kusumoto, 2010), (2) F1, F2, and F3 were determined by entering the LTASS to a pitch detection algorithm (Amano-Kusumoto, 2010), (3) EO1 was determined by calculating the spectral power of the LTASS area over 1 kHz, (4) SB was determined by calculating the ratio of EU1 and EO1 (EU1/EO1), (5) FRT12 was determined by calculating