Signal Processing 139 (2017) 165–172

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Multichannel speech reinforcement based on binaural unmasking

Junhyeong Pak[a], Inyong Choi[b], Yu Gwang Jin[c], Jong Won Shin[a,*]

[a] School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, Korea
[b] Department of Communication Sciences and Disorders, University of Iowa, Iowa City, IA 52242, USA
[c] Corporate R&D Center, SK Telecom Co., Ltd., Seoul 04539, Korea

ARTICLE INFO

Article history:
Received 28 October 2016
Revised 19 April 2017
Accepted 28 April 2017
Available online 2 May 2017

Keywords:
Speech reinforcement
Partial masking effect
Loudness perception
Binaural unmasking
Direction-of-arrival

ABSTRACT

Speech reinforcement or near-end listening enhancement is a technique that modifies the far-end signal to mitigate the effect of the near-end noise, usually based on the power spectra of the far-end signal and the near-end noise. Psychoacoustic experiments have shown that the location of a noise source with respect to that of a signal source affects the amount of masking. Since conventional speech reinforcement methods obtain spectral gain based only on the power spectra, this psychoacoustic phenomenon called binaural unmasking has not been considered in those approaches. In this paper, we propose a novel speech reinforcement algorithm that modifies the far-end speech signal based on both the power spectrum and the direction-of-arrival (DoA) of the noise. Specifically, we have computed the equivalent frontal noise level from the observed noise level and the estimated DoA, and used it to compute spectral gains as in conventional partial loudness restoration-based speech reinforcement. Experimental results showed that the proposed method outperformed the conventional methods based on partial loudness restoration and speech intelligibility index (SII) optimization in terms of the overall perceived quality through subjective listening tests.

1. Introduction

Perceived quality of speech degrades in the presence of ambient noise. The speech enhancement module equipped at the transmitter of the speech communication system reduces the effect of near-end noise on a far-end listener [1–12]. However, the effect of near-end noise on a near-end listener cannot be mitigated by this transmitter-side speech enhancement because the near-end noise arrives directly at the near-end listener's ears. One way to alleviate this problem is to adjust the received far-end monaural signal usually based on estimated near-end noise statistics so that it can be heard more clearly. This approach is called speech reinforcement [13,14], near-end listening enhancement [15–18] or automatic volume control and equalization [19]. Fig. 1 shows the placement of speech enhancement and reinforcement modules in the speech communication system. Various approaches have been proposed for this task including the maintenance of the minimum signal-to-noise ratio (SNR) [15,19], optimization of the speech intelligibility index (SII) [16–18], and restoration of the perceived loudness of speech [13,20]. Some of the approaches assume a strict power constraint that does not allow increasing the power of the signal [16,17]. This constraint is useful for some scenarios, but not appropriate for other applications because the desired far-end speech level in quiet environment is normally much lower than the maximum allowed level.

Most of the speech reinforcement algorithms utilize a single noise statistic only, which is the power spectrum. However, near-end noise arrives both ears, and the relative location of the noise source affects the amount of masking for binaural hearing [21–24]. For example, noises originating from the same direction as the speech signal interfere more than noises from a significantly different direction [21,24–27]. This psychoacoustic phenomenon, called binaural unmasking, implies that the direction-of-arrival (DoA) information of the noise is required in addition to the estimated power spectrum of the noise to compute the optimal reinforcement gain.

In this paper, we propose a multichannel speech reinforcement algorithm that exploits spatial information. The algorithm in [13] is extended to incorporate not only the power spectrum but also the DoA of the noise estimated from dual microphones in a mobile phone. The level of frontal noise that results in a similar amount of masking with the current noise is computed taking binaural unmasking into consideration. It is then used to compute the spectral gains in the same way as in [13]. Experimental results showed that the proposed method was preferred to the algorithm in [13] and

* Corresponding author.
E-mail address: jwshin@gist.ac.kr (J.W. Shin).

http://dx.doi.org/10.1016/j.sigpro.2017.04.021
0165-1684/

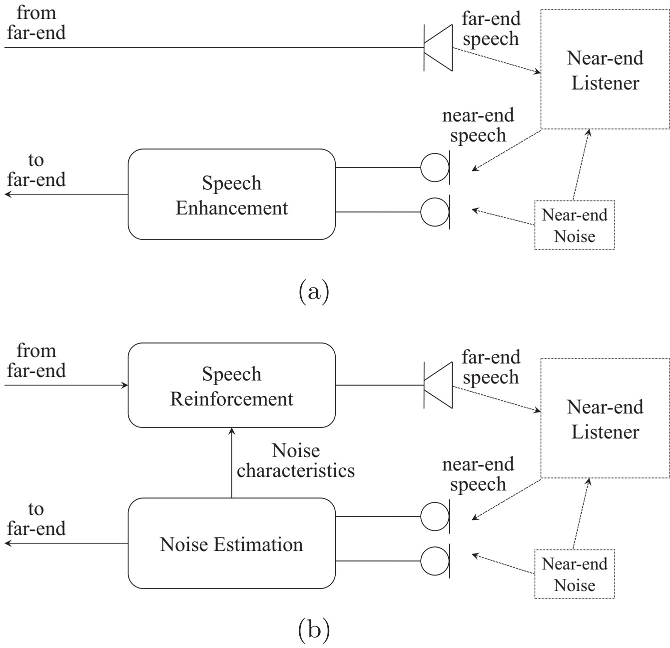

Fig. 1. Block diagram of communication system with speech enhancement (a) and speech reinforcement scheme (b).

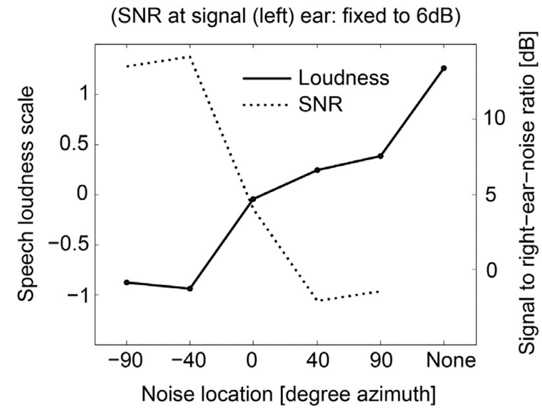

Fig. 2. The loudness scales and the signal-to-right-ear-noise level depending on noise locations.

the SII optimization-based method in [16] through subjective listening tests.

2. Binaural unmasking

A target signal originating from a given direction is less masked by a noise from a different direction than by a noise from the same direction. This is called binaural unmasking or binaural release from masking [21–24]. While this phenomenon has been studied for a few decades, a number of studies have proposed quantitative models of binaural unmasking [24–26,28,29]. However, to the best of our knowledge, none of those models quantifies the masking release for monaural speech with binaural broadband noises. To prove the concept of binaural unmasking in the specific situation where the target speech is presented to the left or right ear only and the maskers come from various directions, we performed a psychophysical experiment. In this experiment, a speech signal was fed to the left ear only, while a background noise coming from various directions was presented binaurally. Using head-related transfer functions (HRTFs, see [27] for details), the direction of the background noise was manipulated into five different azimuth angles from the left (i.e., the speech side, -90°) via the center (front, 0°) to the right (the opposite side, $+90^\circ$). Because we fixed the speech level and the SNR at the left ear, the level of the right-ear noise was increased as the direction of the noise moved rightward due to the head shadow. However, if the effect of binaural unmasking overrides the increase in the masker level, the perceived loudness of the speech signal would be bigger when the noise source is located at the opposite side.

The speech signal was an English sentence of 2.7 s selected from ITU-T P.501 database [30]. The background noise was a restaurant noise selected from the same database. The speech and noise were recorded monaurally at 48 kHz sampling rate, but the speech signal was low-pass filtered up to 7500 Hz to simulate the pass-band of mobile phones. The background noise was not filtered but convolved with HRTFs [27] of five different azimuth angles (-90°, -40°, 0°, $+40^\circ$, and $+90^\circ$). The noise signal at the left-ear was then attenuated by 6 dB, to simulate the reduction of the level by a handset. The noise levels for both ears were again adjusted to yield the fixed SNR of -6 dB at the left ear for all five noise directions. Consequently, the signal-to-right-ear noise ratio varied from 14.1 dB (when the noise was located at -40°) to -2.1 dB (when the noise was located at $+40^\circ$; see the dashed line in Fig. 2). As a result, five binaural signals were created (the fixed-level monaural speech at the left-ear mixed with the binaural background noise of five different directions). A clean monaural speech signal was also used for comparison.

Ten volunteers (seven males and three females) aged 21–42 participated in the experiments. The subjects performed pairwise comparisons; in each trial, we presented a pair of binaural signals (i.e., speech + noise) randomly chosen from all five binaural signals plus one clean monaural speech, without repetition. The subjects were asked to judge in which of the two conditions the speech was perceived to be louder. Because the physical level of the speech signal was fixed for all conditions, the perceived loudness was only affected by partial masking due to the background noise.

The relative loudness of the speech signal in each condition was estimated from pairwise comparisons using Thurstone's case V model [31–33]. Thurstone's loudness scale for one class is the difference of the mean of the class and the average of the means from other classes, assuming that the scores in each class follow Gaussian distribution with a variance of 1/2. The results are given in Fig. 2. The loudness of the speech signal was rated higher when the background noise was located at the opposite side, although the masker energy at the right ear was even higher in those conditions. This result empirically proves that the binaural unmasking is strong (enough to override more than 10 dB increase in the opposite-ear masker level) and must be considered during speech reinforcement in cases where the background noise has a clear direction.

3. Speech reinforcement based on binaural unmasking

Perceived loudness of a speech signal diminishes in the presence of background noise. Based on this phenomenon, the speech reinforcement based on the loudness restoration of the speech signal was proposed in [13]. Spectral gains that make the perceived loudness of the reinforced speech in noise for each band similar to that of the original far-end speech in clean condition are computed by adopting the loudness perception model in [34]. Essentially, the gains become functions of the excitations at cochlea incurred by speech and noise, which are filtered and frequency-warped version of the power spectrum of signal and noise, respectively.

Conventional speech reinforcement algorithms including [13] utilize only the estimated power spectrum of the far-end signal and the near-end noise. However, as we have seen in the