Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 964–974

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Speech enhancement based on soft audible noise masking and noise power estimation ☆

Rongshan Yu *

Department of Signal Processing, Institute for Infocomm Research (I2R), 1 Fusionopolis Way, #21-01 Connexis (South Tower), Singapore

Received 28 December 2012; received in revised form 20 May 2013; accepted 22 May 2013
Available online 25 June 2013

Abstract

This paper presents a perceptual model based speech enhancement algorithm. The proposed algorithm measures the amount of the audible noise in the input noisy speech based on estimation of short-time spectral power of noise signal, and masking threshold calculated from the estimated spectrum of clean speech. An appropriate amount of noise reduction is chosen based on the result to achieve good noise suppression without introducing significant distortion to the clean speech. To mitigate the problem of "musical noise", the amount of noise reduction is linked directly to the estimation of short-term noise spectral amplitude instead of noise variance so that the spectral peaks of noise can be better suppressed. Good performance of the proposed speech enhancement system is confirmed through objective and subjective tests.

Keywords: Speech enhancement; Speech processing; Auditory model; Perceptual model; Noise estimation; Noise suppression; Noise tracking

1. Introduction

Speech enhancement technology has been widely used in telecommunication systems to improve the quality of voice communication in noisy environments. Usually it is performed directly on the output of the microphone in the transmission side to achieve the best speech enhancement quality. However it is possible to apply the speech enhancement system inside the telecommunication network or use it in the playback device. For economic reasons most systems are single microphone based solutions where the speech enhancement is done on the output of a single microphone, although better speech enhancement results can be achieved by using a microphone array system with more than one microphone.

In principle, a single microphone speech enhancement system uses some sort of adaptive filtering operation to attenuate the time/frequency (T/F) regions of the noisy speech signal that have low Signal-to-Noise-Ratios (SNR), and preserve those with high SNR. By doing so, the essential parts of the speech signal is thus preserved while the noise level is greatly reduced, leading to an enhanced signal with reduced noise level. Various speech enhancement systems along this line have been proposed in the literature, e.g., spectral subtraction (Boll, 1979), Wiener filter (Widrow and Stearns, 1985), MMSE-STSA (Ephraim and Malah, 1984), and MMSE-LSA (Ephraim and Malah, 1985). In these algorithms, some attenuation rules are used to decide which T/F regions of the noisy speech should be attenuated and by how much. Usually, these attenuation rules are optimized in such a way that the enhanced speech is as close as possible to the clean speech signal in the input; and the difference among these attenuation rules mainly result from different statistical models of the signals assumed as well as different distortion measurements used in the optimization.

☆ This research was done while the author was a staff engineer with Dolby Laboratories, Inc., San Francisco, CA 94103, USA.

* Tel.: +65 64082629.
E-mail addresses: ryu@i2r.a-star.edu.sg, rongshanyu@ieee.org.

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.05.006

Clearly, the quality of a single-microphone speech enhancement system described above is to a large extent determined by the suppression rule it uses. In general, a suppression rule with stronger attenuation will lead to less noisy output; however, the speech signal will become more distorted. Conversely, a suppression rule with more moderate attenuation will produce less distorted speech signal while it can only achieve limited amount of noise reduction. For this reason, a careful trade-off has to be made to balance the amount of the noise suppression with the speech distortion for optimal quality. To this end, auditory masking model, which has been successfully applied in wideband audio coding (Johnston, 1988), has recently been introduced in speech enhancement systems (Gustafsson et al., 1998; Lin et al., 2003; Virag, 1999; Tsoukalas et al., 1997; Hu and Loizou, 2004; Jabloun and Champagne, 2003; Hansen et al., 2006). In these systems, the masking properties of speech signal are employed to identify the perceptual significant noise components of the noisy speech signal, which are then subtracted from the speech signal for noise reduction. Various heuristics have been proposed in this system to incorporate the masking threshold of speech signal into the subtraction equations in order to obtain a trade-off between audible noise suppression and speech distortion.

In this paper we propose a perceptual model based speech enhancement algorithm. The proposed algorithm achieves good noise suppression qualities by using an attenuation rule that carefully balances the amount of reduction of the audible noise and the amount of distortion introduced to the clean speech. To this end, short-term spectral amplitudes of both the clean speech and noise signals are estimated continuously in the algorithm. Masking threshold of the estimated clean speech amplitude is then calculated by using a perceptual model. After that, the amount of audible noise in the noisy signal is calculated by contrasting the estimated noise amplitude to the masking threshold and an appropriate amount of attenuation is chosen based on a soft audible noise suppression principle that minimizes a cost function that explicitly includes the amount of audible noise and speech distortion in the enhanced speech signal. Since the auditory masking effect is only a short time phenomenon with a limited duration (Johnston, 1988), in the proposed algorithm the amount of attenuation of the proposed algorithm is linked directly to the estimation of the short-term noise spectral amplitudes instead of long-term noise variance. As a result, it provides superior suppression of the noise peaks in the frequency domain, leading to fewer "musical noise" artifacts (Cappe, 1994) in the enhanced speech signal.

Noise variance estimation plays an important role in determining the quality of a speech enhancement system, particularly in an environment with non-stationary noise. One popular choice of the noise variance estimator in both research literature and commercial speech enhancement implementations is the Voice Activity Detection (VAD) based approach, where the noise estimation is updated only when speech is not present in the input. The performance of the VAD approach strongly depends on the accuracy of the voice detection, which is a difficult task in particular for signals with low SNR. In addition, this method precludes the possibility of updating the noise estimation when the speech signal is present, which is inefficient since there may still be spectral bands where the speech level is weak even during speech segments. Another widely cited method is the Minimum Statistics (MS) noise estimator (Martin, 2001). In principle, the MS method keeps a record of historical samples for each spectral location, and the noise level is estimated based on the minimum signal level from the record. It is reported that the MS method achieves good tracking performance for non-stationary noises; however, it has a high memory demand and is not applicable to devices with limited memory resources.

To address these issues, in this paper we adopted a low-complexity noise variance estimation algorithm previously described in Yu (2009). In this algorithm, the instantaneous noise power is estimated each frame based on information from the incoming signal and the current estimated distribution parameters. After that, the distribution parameters, including the noise variance that we are interested in, are refined from the expectation results. Instead of using a trained gain function for noise power estimation as proposed in Erkelens and Heusdens (2008), naive minimum mean-square-error (MMSE) noise power expectation is used and the potential estimation bias problem is addressed by using a bias estimation correction method. The proposed algorithm has very low computational power and memory requirements while still delivering satisfactory performance for various noise types in our tests.

The rest of this paper is organized as follows. In Section 2, the principles of proposed soft audible noise suppression algorithms are described. Implementation issues of the proposed algorithm, including noise amplitude and variance estimation and the masking threshold calculation, are presented in detail in Section 3. The performance of the each of the proposed algorithms is evaluated in Section 4. Finally, Section 5 presents the conclusions drawn from this work.

2. Principle

2.1. Signal model

We consider the following additive signal model for a noisy speech signal:

$$\mathbf{Y}_k(m) = \mathbf{X}_k(m) + \mathbf{D}_k(m), \quad k = 1, \ldots, K, \tag{1}$$

where $\mathbf{Y}_k(m), \mathbf{X}_k(m)$, and $\mathbf{D}_k(m)$ are complex-valued short-time Fourier Transform (STFT) coefficients of the noisy speech signal $y(n)$, clean speech signal $x(n)$, and noise signal $d(n)$ respectively. Here k is the subband index, K is the total number of subbands, and m is the frame index.

In most current speech enhancement algorithms the speech and the noise signals are usually modeled as