Applied Acoustics 74 (2013) 770–781

Contents lists available at SciVerse ScienceDirect

Applied Acoustics

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

Noise estimation based on time–frequency correlation for speech enhancement

Wenhao Yuan, Jiajun Lin *, Wei An, Yu Wang, Ning Chen

School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

ARTICLE INFO

Article history:
Received 16 May 2012
Received in revised form 16 October 2012
Accepted 20 November 2012
Available online 8 January 2013

Keywords:
Noise estimation
Speech enhancement
Minimum search
Improved Minima Controlled Recursive Averaging

ABSTRACT

As a fundamental part of speech enhancement, noise estimation is particularly challenging in highly non-stationary noise environments. In this work, we propose an effective algorithm on the basis of the "Improved Minima Controlled Recursive Averaging (IMCRA)" with the objective to improve the performance of noise estimation. The main contributions of this work are: (i) in the algorithm, a rough decision about speech presence is proposed by calculating the autocorrelation and cross-channel correlation of the T–F (Time–Frequency) units; (ii) with this decision, we refine the smoothing parameters for the smoothing of noisy power spectrum and the recursive averaging in noise spectrum estimation as well as the weighting factor for the *a priori* SNR (Signal to Noise Ratio) estimation in the IMCRA; (iii) we improve the search of local minima during spectral bursts by adding a minimum search with a shorter window. Extensive experiments are carried out to evaluate the performance of our proposed algorithm. The experimental results illustrate that, compared with the IMCRA, the proposed approach significantly improves the accuracy of noise spectrum estimation and the quality of enhanced speech in the typical noise situations.

1. Introduction

Due to universal applicability and simplicity, single channel speech enhancement has been being a hot research spot of speech enhancement, for several years, that is an indispensable step in various fields, such as speech communication, speech coding and speech recognition in noisy environments [1–7]. As being a necessary step in most single channel speech enhancement algorithms, noise estimation significantly affects the performance of speech enhancement. Traditional noise estimation methods utilize voice activity detectors (VADs) for detecting the presence and absence of speech in noisy speech, and update the noise estimation during speech absence [8]. The VADs can achieve good results in stationary noise environments (e.g., white noise) because they mainly take advantage of the energy statistical properties and other characteristics of speech and noise signals. However, in real noise environments (e.g., factory noise), considering the rapid varying of energy statistical properties of the noise, VADs usually do not work well and result in losing track of the immediate changes of noise. Therefore, an urgent research work is to develop a more accurate and robust noise estimation algorithm such that it can update the noise power spectrum continuously without relying on the judgments of VADs.

Recently increasing efforts can be found in literature that focuses on noise estimation. Doblinger [9] updated the noise estimate continuously through tracking the spectral minima frame by frame. However, this method may tend to attenuate the speech spectrum, because it fails to differentiate between an increase in noise floor and an increase in the speech spectrum level. Hirsch and Ehrlicher [10] proposed a simple noise estimation method by using a first order recursive averaging, which updates the noise estimate by comparing the power spectrum of current frame to the noise estimate of past frames. However, this method fails to update the noise estimate in the case of abrupt increase of noise floor. Ris and Dupont [11] estimated the noise by combining the above techniques with narrow-band spectral analysis, while Stahl et al. [12] proposed a quantile-based noise estimation algorithm which filters out speech peaks and estimates the noise from the remaining spectrum with the use of a non-linear filter. However, the above two methods may fail to adapt fast to highly-varying noise. Martin [13] proposed a noise estimation algorithm, based on minimum statistics (MS), which tracks the minima values of a smoothed spectrum of the noisy speech over a finite window, and then multiplies the result by a bias factor to achieve the unbiased estimate of noise spectrum. The major drawback of this method is that the update of the noise spectrum spends more time than the duration of the minimum-search window when the noise floor increases abruptly. Cohen and Berdugo [14] proposed a minima controlled recursive averaging algorithm (MCRA). The MCRA searches the local minimum similarly to MS, and then compares the ratio of the noisy speech to the local minimum against a threshold to find the noise-only regions.

* Corresponding author. Tel./fax: +86 021 64253511.
E-mail addresses: ywhecust@126.com (W. Yuan), jjlin@ecust.edu.cn (J. Lin), anweiunl@gmail.com (W. An), rain198689@163.com (Y. Wang), nchen@ecust.edu.cn (N. Chen).

0003-682X/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.apacoust.2012.11.007

The noise estimate is updated by tracking the noise-only regions of the noisy speech spectrum. In [15], the improved MCRA (IMCRA) approach was proposed, which exploits a different method to track the noise-only regions based on the estimated speech presence probability. However, both the MCRA and IMCRA might take twice as much of the duration of the search window to adapt to a noise burst. Rangachari et al. [16] described a noise estimation algorithm based on voice activity detection, which updates the noise estimate in each frame. In [17], the method was further improved by the way of making the speech presence decision based on the ratio of noisy speech spectrum to its local minimum against a frequency-dependent threshold. This method reduces adaptation time to noise bursts, but it may occasionally attenuate some speech components.

In sum, the main drawbacks of these noise estimation algorithms are the inaccuracies for distinguishing speech from noise and the delayed responses to the abrupt increases of noise floor. In this work, we propose a novel noise estimation algorithm based on the IMCRA so that it can improve the tracking ability and the accuracy of noise estimate by introducing two feature functions that can effectively reflect the correlation of the signals.

(i) The T–F correlation is a significant feature to differentiate speech from noise. The IMCRA takes into account the correlation between consecutive frames and adjacent frequencies by carrying out the smoothing of noisy power spectrum, the recursive averaging in the noise spectrum estimation and the non-linear recursive procedure in the *a priori* SNR estimation. However, in the calculation of the speech presence probability based on the ratio of noisy speech spectrum to the local minima, the IMCRA mainly focuses on the comparison of the power spectrum in different frames, and does not entirely consider the relation between correlation features and speech presence.
In this paper, we first adopt the peripheral auditory processing in the CASA (Computational Auditory Scene Analysis) [18] to make an initial processing of the noisy speech signal, and convert the signal to T–F units distributing in 128 channels. Then, the autocorrelation and the cross-channel correlation of the T–F units are calculated, which can reflect the coherent relationship among different frames and different frequency bins. By comparing the calculation of two correlation functions to fixed thresholds, a rough decision about speech presence is made.

(ii) The smoothing of noisy power spectrum and the update of noise spectrum estimation has great impact on the resultant noise estimate. In the IMCRA, to implement the smoothing and the update, some smoothing parameters are introduced and defined as constants. However, the optimal values of the parameters should vary according to the speech presence probability rather than be certain constants. In addition, the estimate of the *a priori* SNR is one of the critical factors to affect the speech enhancement. In the IMCRA, the calculation of the speech presence probability used in noise estimation depends on the estimate of the *a priori* SNR, therefore the estimation of the *a priori* SNR is also one of the critical factors to affect the noise estimate. The weighting factor for the *a priori* SNR estimation in the IMCRA is a constant.
From the above discussion, it is worthy to be mentioned that that the choice of these parameters is closely related to the accuracy of the noise estimate. This paper refines above parameters to some values depending on the speech presence of each frame based on the rough decision we made previously.

(iii) To reduce the delay of minimum search in the case of spectral bursts, we add a parallel minimum search that is similar to the original one but using a shorter length window. And we combine it with the obtained rough decision about speech presence to improve the minimum search when spectral bursts are detected.
The rest of this paper is organized as follows. Section 2 briefly reviews the IMCRA method, and Section 3 presents the rough decision about speech presence based on correlation features. In Section 4, the decision is used to adjust the smoothing parameters for the smoothing of the noisy power spectrum and the recursive averaging in the noise spectrum estimation, and also the weighting factor for the *a priori* SNR estimation. Section 5 introduces a new minimum search method by combining the decision with two parallel minimum searches. Section 6 evaluates the performance of the proposed algorithm compared to the IMCRA. Finally, the conclusions are given in Section 7.

2. Review of improved minima controlled recursive averaging

Let y denote an observed noisy signal in the time domain, which is the sum of a clean speech x and an uncorrelated additive noise d. By applying the short-time Fourier transform (STFT), we have

$$Y(k,l) = X(k,l) + D(k,l) \tag{1}$$

in the time–frequency domain, where k represents the frequency bin index, and l is the frame index.

In the IMCRA, the noise is estimated by recursively averaging past spectral power values of the noisy measurement during periods of speech absence and holding the estimate during speech presence [15]. Under speech presence uncertainty, the conditional speech presence probability is employed, and the recursive averaging can be obtained by

$$\bar{\lambda}_d(k,l+1) = \tilde{\alpha}_d(k,l)\bar{\lambda}_d(k,l) + [1-\tilde{\alpha}_d(k,l)]|Y(k,l)|^2 \tag{2}$$

where

$$\tilde{\alpha}_d(k,l) \triangleq \alpha_d + (1-\alpha_d)p(k,l) \tag{3}$$

is a time-varying frequency-dependent smoothing parameter. $\alpha_d (0 < \alpha_d < 1)$ denotes a smoothing parameter, and $p(k,l)$ is the conditional speech presence probability. Through introducing a bias compensation factor β, the noise estimate is given by

$$\hat{\lambda}_d(k,l+1) = \beta \cdot \bar{\lambda}(k,l+1) \tag{4}$$

In order to calculate the speech presence probability, two iterations of smoothing and minimum tracking are carried out. The time smoothing in the first iteration is performed by a first-order recursive averaging as

$$S(k,l) = \alpha_s S(k,l-1) + (1-\alpha_s)S_f(k,l) \tag{5}$$

where α_s $(0 < \alpha_s < 1)$ is a smoothing parameter, and $S_f(k,l)$ is obtained by the frequency smoothing of the noisy power spectrum

$$S_f(k,l) = \sum_{i=-\omega}^{\omega} b(i)|Y(k-i,l)|^2 \tag{6}$$

where b denotes a normalized window function. The time smoothing in the second iteration is similar to that in the first iteration, and utilizes the same smoothing parameter.

The method of minimum search in the IMCRA is in accordance with that used in MS [13], the local minima of $S(k,l)$ is searched within a finite window of length D, for each frequency bin

$$S_{\min}(k,l) \triangleq \min\{S(k,l') | l-D+1 \leqslant l' \leqslant l\}. \tag{7}$$

To reduce the computational complexity, the window of D samples is generally divided into U sub-windows of V samples ($D = UV$).