Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 79 (2016) 30–46

www.elsevier.com/locate/specom

Speech enhancement based on AR model parameters estimation

Feng Deng, Changchun Bao*

Speech and Audio Signal Processing Lab, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China

Received 25 May 2015; received in revised form 22 February 2016; accepted 23 February 2016
Available online 3 March 2016

Abstract

In this paper, we propose a speech and noise auto-regressive (AR) model parameters estimation method under noisy conditions used for speech enhancement, which exploits a priori information about speech and noise spectral shapes (parameterized as AR coefficients) described by trained codebooks. The expectation maximization (EM) algorithm is first employed to obtain AR gains of speech and noise, which correspond to each pair of codebook entries of speech and noise spectral shapes. Then the K-nearest neighbor (KNN) rule is used to select some candidates from the optimized AR parameters (AR coefficients and AR gains) of speech and noise for constructing the weighted Wiener filter (WWF). Furthermore, by using sigmoid function, we propose a posteriori speech-presence probability (SPP) estimation method. Combining the a posteriori SPP with the WWF, the residual noise of enhanced speech is effectively reduced. The test results demonstrate the performance superiority of the proposed speech enhancement scheme compared to the reference methods.

Keywords: Speech enhancement; AR model; Expectation maximization; Spectral shapes codebook; Wiener filter.

1. Introduction

Conventional single-channel speech enhancement algorithms, such as Wiener filtering (Loizou, 2007), spectral subtraction (Boll, 1979) and Bayesian short-time spectral amplitude (STSA) estimation (Ephraim and Malah, 1984; Ephraim and Malah, 1985; You et al., 2005; Xie and Zhang, 2014), have a common feature that they have no a priori information about speech and noise, which limits the performance of speech enhancement.

To address speech enhancement by using a priori information about speech and noise, the trained codebooks that are composed of auto-regressive (AR) model coefficients have been used to exploit a priori information about speech and noise for speech enhancement (Srinivasan et al., 2003 , 2006 , 2007; Rosenkranz, 2010; Rosenkranz and Puder, 2012), where the AR coefficients are used to describe the spectral shapes of speech and noise. In Srinivasan et al. (2003 , 2006), the maximum likelihood (ML) method was used to obtain an optimal speech and noise AR coefficients from respective codebooks, and the corresponding AR gains of speech and noise were estimated as well. Here, we refer to the AR coefficients and the corresponding gains as AR parameters. Subsequently, the Wiener filter was constructed for enhancing the noisy speech based on the optimal AR parameters of speech and noise. Unlike the ML scheme treated the AR parameters as deterministic parameters, the Bayesian minimum mean squared error (MMSE)-based schemes were presented (Srinivasan et al., 2007; Rosenkranz, 2010; Rosenkranz and Puder, 2012), in which the AR parameters were considered as the random variables, and thus the AR parameters were estimated as a weighted sum of all codebook entries of speech and noise spectral shapes and the corresponding AR gains estimated by ML method.

For the ML-based AR gains estimation given in Srinivasan et al. (2003, 2006, 2007), Rosenkranz (2010) and Rosenkranz and Puder (2012), the solutions were approximated by minimizing the Itakura–Saito distortion (Itakura and Saito, 1970) between the observed and estimated noisy spectra instead of maximizing the log-likelihood. However, this kind of approximation is often not accurate due to the steepness of likelihood function, which results in inaccurate AR gains estimation of speech and noise. In addition, as a result of the Bayesian methods (Srinivasan et al., 2007; Rosenkranz, 2010; Rosenkranz and Puder, 2012) used to obtain the AR

* Corresponding author. Tel.: +86 10 67391635; fax: +86 10 67391625.
E-mail address: baochch@bjut.edu.cn (C. Bao).

http://dx.doi.org/10.1016/j.specom.2016.02.006
0167-6393/© 2016 Elsevier B.V. All rights reserved.

parameters estimation about speech and noise, the linear combinations (i.e. weighted sum) of all codebook entries of speech and noise spectral shapes are generally used to describe the spectral shape of the observed noisy speech. However, the combinations of speech spectra are not restricted to look like speech. For example, the linear combination (i.e. weighted sum) of two spectra with three formants may contain six formants that are not possible for general utterance of human beings. Such combined speech spectrum may represent a noise-like observation that we do not expect. Similarly, noise spectrum is often combined to form an unexpected speech-like spectrum. In other words, the combinations of speech spectral shapes and noise spectral shapes sometimes result in an ambiguity of spectral shapes of speech and noise. As a result, the separation of the speech and noise components for speech enhancement becomes difficult. Here, we refer to such problem as the *ambiguity problem* which increases with the codebook sizes of speech and noise spectral shapes. Thus, if the codebook sizes of spectral shapes of speech and noise are relatively small, the ambiguity problem is less severe. However, decreasing the codebook sizes of spectral shapes of speech and noise, the speech enhancement performance would be decreased. To avoid ambiguity problems, it is natural to restrict the number of codebook entries for the combinations of speech spectral shapes and noise spectral shapes. Although the ML scheme in Srinivasan et al. (2003 , 2006) can avoid such ambiguity problem to some extent, it only selects one pair of codebook entries of spectral shapes of speech and noise, which is too crude to perform well for speech enhancement. Moreover, to reduce the ambiguity problem, they trained a set of codebooks with smaller size for noise spectral shape (each codebook describes a particular noise type) and then the noise classification was required to choose an active codebook. However, these small codebooks of noise spectral shapes cannot describe all the noise scenarios in the real world. As a result, their robustness to adapt the real noise environments is poor. So the inaccurate AR gains estimation of speech and noise and the ambiguity problem limit the overall performance of conventional codebook-based methods.

In this paper, we propose a new solution for the aforementioned problems: firstly, the expectation maximization (EM) algorithm (Bilmes, 1997; Ephraim, 1992) is applied to obtain the accurate AR gains of speech and noise corresponding to each pair codebook entries of spectral shapes of speech and noise, which can provide more accurate a priori information about speech and noise for speech enhancement. Secondly, the K-nearest neighbor (KNN) rule (Zhang and Yang, 2010) is employed to select some candidates from the optimized speech and noise AR parameters which are more possibly produced by the observed noisy speech. Finally, they are used to construct the weighted Wiener filter (WWF) for enhancing noisy speech. In this way, only a few codebook entries of spectral shapes of speech and noise and the corresponding gains have a significant contribution to the WWF for the observed noisy speech. Therefore, the spectral ambiguity problems are reduced, and then we are allowed to train a large, general noise codebook that covers most possible noise scenarios. Thus we do not require a noise classification. The robustness of environment adaptation of the proposed method is improved as well as the enhancement performance.

In addition to the aforementioned two problems, the conventional codebook-based methods (i.e. ML and Bayesian MMSE) in Srinivasan et al. (2003 , 2006 , 2007), Rosenkranz (2010) and Rosenkranz and Puder (2012) have another inherent problem: the spectral shape codebook of speech could not model the spectral fine structure of voiced speech well. Therefore, the clearly audible noise is still remained in the voiced segments of the estimated speech, which reduces the perceptual quality of enhancement system. In order to improve the perceptual quality of the estimated speech, we propose a posteriori speech-presence probability (SPP) (Malah et al., 1999; Deng et al., 2014; Fu and Wang, 2010) estimation method by using sigmoid function. And the a posteriori SPP combined with the WWF can remove the residual noise effectively. Experimental results demonstrate that the proposed method outperforms the reference methods.

The remainder of this paper is organized as follows. In Section 2, the conventional ML parameters estimation method is described. In Section 3, we present the proposed AR parameters estimation method for speech enhancement. The performance evaluation is presented in Section 4, and Section 5 gives the conclusions.

2. Conventional ML method

In this section, we first present the signal model of the AR parameters estimation for speech enhancement. Then the conventional ML method for AR parameters estimation is introduced. Meanwhile, its drawbacks are analyzed.

2.1. Signal model

Assuming the clean speech $\boldsymbol{x}_n$ is contaminated by an additive noise $\boldsymbol{w}_n$, where the speech and noise are statistically independent, and then we have the noisy speech

$$\boldsymbol{y}_n = \boldsymbol{x}_n + \boldsymbol{w}_n \tag{1}$$

where $\boldsymbol{y}_n$, $\boldsymbol{x}_n$ and $\boldsymbol{w}_n$ are the N-dimension vectors of noisy speech, clean speech and noise, respectively. n is the frame index and we ignore it for convenience later.

Under the Gaussian distribution assumption, the probability density function (pdf) of the clean speech given the AR parameters can be modeled as Srinivasan et al. (2006):

$$p(\boldsymbol{x}|\boldsymbol{\alpha}_x, g_x) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{R}_x|^{1/2}} \exp(-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{R}_x{}^{-1}\boldsymbol{x}) \tag{2}$$

where $\boldsymbol{\alpha}_x$=$[1\ \alpha_{x1}\ \dots\ \alpha_{xp}]^{\mathrm{T}}$ is the vector of AR coefficients of speech with order p, $\boldsymbol{R}_x$=$g_x(\boldsymbol{A}_x^T\ \boldsymbol{A}_x)^{-1}$ is the covariance matrix where $\boldsymbol{A}_x$ is the $N\times N$ lower triangular Toeplitz matrix with $[1\ \alpha_{x1}\ \dots\ \alpha_{xp}\ 0\dots0]^{\mathrm{T}}$ as the first column and the g_x is the AR gain of speech.

Similarly, the noise's pdf is modeled as well, where the covariance matrix $\boldsymbol{R}_w$ of noise is defined analogous to $\boldsymbol{R}_x$. Therefore, combining speech and noise models, we can obtain