A Data-Driven Speech Enhancement Method Based on Modeled Long-Range Temporal Dynamics

Yue Hao, Changchun Bao, Feng Bao, Feng Deng

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

s201302071@emails.bjut.edu.cn, baochch@bjut.edu.cn, {baofeng,dengfeng}@emails.bjut.edu.cn

10.21437/Interspeech.2015-415

Abstract

In this paper, a data-driven speech enhancement method based on modeled long-range temporal dynamics (LRTDs) is proposed. First, given speech and noise corpora, Gaussian Mixture Models (GMMs) of the speech and noise can be trained respectively based on the expectation-maximization (EM) algorithm. Then, the LRTDs are obtained from the GMM models. Next, based on the LRTDs, a noise robustness longest segment searching (NRLSS) method combined with the Vector Taylor Series (VTS) approximation algorithm is adopted to search the longest matching speech and noise segments (LMSNS) from speech and noise corpora. Finally, using the obtained LMSNS, the estimation of speech spectrum is achieved. Furthermore, a modified Wiener filter is constructed to further eliminate residual noise. The test results show that the proposed method outperforms the state-of-the-art speech enhancement methods.

Index Terms: speech enhancement, LRTDs, GMM, NRLSS, VTS, modified Wiener filter

1. Introduction

The goal of speech enhancement is to remove noise from noisy speech for improving speech quality and intelligibility. Currently, single-channel speech enhancement methods typically consist of two classes: unsupervised methods and supervised methods. For the unsupervised techniques, such as Wiener filter (WF) method [1], spectral subtraction (SS) method [2], minimum mean-square error (MMSE) method [3], weighted Euclidean distortion measure (WEDM) method [4], a common problem is that there is always a trade-off between noise suppression and speech distortion. By contrast, for the supervised methods, such as codebook-based (CB) method [5] and nonnegative matrix factorization (NMF) method [6], can achieve remarkable performance in noise suppression and speech enhancement, since they provided more priori information about speech and noise signal. However, these methods failed to capture the inter-frame dependency of speech signal. Thus, they often suffer poor performance under non-stationary noise environment because of their weak predictability for the fast-varying noise.

In recent years, researches have emphasized on cross-frame importance for improving speech quality in highly non-stationary noise conditions, such as hidden Markov model (HMM) methods [7, 8, 9]. However, their state dynamics are unrealistic for representing temporal dynamics of speech in a long-range period under the first-order Markov chain assumption. The HMM-based method was further developed by using a data-driven method given in [10], in which a corrupted signal was reconstructed as a new clean signal from a large speech corpus. Although the data-driven method improved the desirable performance in producing better quality and natural sounding output, it could not explicitly model the long-range temporal dynamics (LRTDs) of the target speech. This implies that it will be difficult to separate speech from noisy speech in a short-term period, due to the non-stationary features of speech and noise.

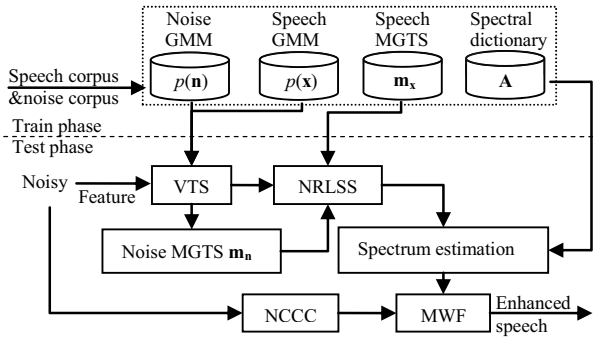

Figure 1: *A block diagram of the proposed method.*

For solving the aforementioned problems, in this paper, we extend the data-driven method used in speech segmentation and recognition [11] and as much as possible to extract LRTDs of speech and noise signals (i.e. GMMs, maximum Gaussian time sequence: MGTS) for improving the performance of speech enhancement. The main contribution of this paper is to improve LRTDs modeling of speech and noise, and applies it to speech enhancement. The LRTDs play an important role in accurately separating out the clean speech and tracking fast-varying noise types from noisy speech. The data-driven framework [11] has also been applied to speech enhancement in [12]. However, there are several differences: noise modeling for speech estimation, VTS expansion to adapt the clean statistics and a modified wiener filter to eliminate background noise. The block diagram of the proposed method is illustrated in Figure 1. In the training phase (top dotted in Figure 1), speech and noise corpora are first prepared. Feature (i.e. Mel-Frequency Cepstral coefficient: MFCC) extraction is conducted on the speech and noise corpora. Using the MFCC features, speech GMM and multiple noise GMM models each for one type of noise can be trained offline based on the expectation-maximization (EM) algorithm [13], respectively. Then the MGTS of speech corpus can be derived from the speech GMM, and the paralleled 'spectral dictionary'is generated. In the test phase, for a given noisy speech segment, a classification is made and a proper

noise GMM model with its MGTS are determined. Using the speech and noise MGTS, the corresponding longest matching speech and noise segments (LMSNS) are identified based on a noise robustness longest segment searching (NRLSS) method combined with VTS [14], which is used widely in robust speech recognition. Based on the determined LMSNS, we estimate the spectrum by concatenating the corresponding 'spectral dictionary'. Moreover, according to the normalized cross-correlation coefficient (NCCC) [15], the modified Wiener filter (MWF) is constructed, which could further eliminate the residual noise during silence or unvoiced segments. The magnitude spectrum of the enhanced speech is obtained by filtering the noisy speech through the MWF.

The remainder of this paper is organized as follows. Section 2 presents an overview of the data-driven framework proposed in [11]. The proposed speech enhancement method based on data-driven framework is described in Section 3. The performance evaluation results are shown in Section 4 and Section 5 gives the conclusions.

2. Overview of data-driven framework

2.1. The LRTDs modeling

In contrast to most conventional modeling methods for speech enhancement, a novel speech corpus modeling method [11] based on data-driven framework is described in this section. This model is good at representing LRTDs with any segment and any length in the training sentences of speech corpus.

Let $\{\mathbf{x}_i : i = 1, 2, ..., I_{\mathbf{x}}\}$ be a complete speech MFCC feature sequence. $I_{\mathbf{x}}$ is the number of speech frames, and $\mathbf{x}_i$ is the MFCC feature vector at frame i. Using vector $\mathbf{x}$ to represent a set including $I_{\mathbf{x}}$ MFCC feature vectors, a speech GMM probability distribution function (PDF) with M mixtures is first described as

$$p(\mathbf{x}) = \sum_{m_{\mathbf{x}}=1}^{M} g_{m_{\mathbf{x}}} N(\mathbf{x}|\mathbf{u}_{m_{\mathbf{x}}}, \mathbf{\Sigma}_{m_{\mathbf{x}}}) \quad (1)$$

where $N(\mathbf{x}|\mathbf{u}_{m_{\mathbf{x}}}, \mathbf{\Sigma}_{m_{\mathbf{x}}})$ indicates the $m_{\mathbf{x}}^{th}$ Gaussian distribution, $g_{m_{\mathbf{x}}}$ is the corresponding mixture weight, $\mathbf{u}_{m_{\mathbf{x}}}$ and $\mathbf{\Sigma}_{m_{\mathbf{x}}}$ are the mean vector and the covariance matrix, respectively.

Based on the speech GMM, a MGTS $\mathbf{m}_{\mathbf{x}}$ on the speech corpus is then derived as

$$\mathbf{m}_{\mathbf{x}} = \{m_{\mathbf{x},i} : i = 1, 2, ..., I_{\mathbf{x}}\} \quad (2)$$

where $m_{\mathbf{x},i}$ represents the index of Gaussian mixture that maximizes the likelihood of Gaussian distributions $\{N(\mathbf{x}|\mathbf{u}_{m_{\mathbf{x}}}, \mathbf{\Sigma}_{m_{\mathbf{x}}}), m_{\mathbf{x}} = 1, ..., M\}$ at frame i.

Finally, a pre-recorded 'spectral dictionary' $\mathbf{A}$ is generated, which is in parallel with $\mathbf{m}_{\mathbf{x}}$

$$\mathbf{A} = \{\mathbf{a}_i : i = 1, 2, ..., I_{\mathbf{x}}\} \quad (3)$$

where $\mathbf{a}_i$ denotes the magnitude spectrum of speech at frame i.

2.2. The longest matching speech segment searching

To search the matched segment more accurately, a longest matching speech segment (LMSS) searching algorithm based on MAP criterion [11] is simply described as follows.

Let the MFCC feature sequence $\mathbf{s}_{t:t+\tau} = \{\mathbf{s}_n : n = t, t+1, \ldots, t+\tau\}$ denotes a test speech segment from frame t to frame $t+\tau$, and the MGTS $\mathbf{m}_{\mathbf{x},u:u+\tau} = \{m_{\mathbf{x},i} : i = u, u+1, \ldots, u+\tau\}$ denotes speech segment taken from (2). For each $\mathbf{s}_{t:t+\tau}$ at frame t, we can find its LMSS $\mathbf{m}^t_{\mathbf{x},u:u+\tau_{\max}}$ by using the MAP criterion[11]

$$\mathbf{m}^t_{\mathbf{x},u:u+\tau_{\max}} = \arg\max_{\tau} \max_{\mathbf{m}_{\mathbf{x},u:u+\tau}} p(\mathbf{m}_{\mathbf{x},u:u+\tau}|\mathbf{s}_{t:t+\tau}) \quad (4)$$

where $p(\mathbf{m}_{\mathbf{x},u:u+\tau}|\mathbf{s}_{t:t+\tau})$ denotes the likelihood function of $\mathbf{m}_{\mathbf{x},u:u+\tau}$ given $\mathbf{s}_{t:t+\tau}$.

In the absence of background noise, given each segment of the test speech, we can accurately obtain its LMSS according to (4). However, for the noise presence condition, if we just apply the above data-driven framework to deal with the problem of speech enhancement, the robustness of noise adaptation will be influenced for searching the correct LMSS from noisy speech. To solve such problem, we propose a NRLSS method. Not only the offline trained speech and noise corpora models are exploited, but also noise information is incorporated to search the LMSNS, which can improve the performance of speech enhancement greatly.

3. Proposed speech enhancement method

In Section 3.1 and 3.2, we aim to search the LMSNS using a NRLSS method combined with VTS algorithm based on the above data-driven framework. Using the LMSNS, we can get the spectral estimation of speech and noise (Section 3.3). The estimated spectra can be used to construct MWF (Section 3.4).

3.1. The NRLSS method with noise GMM model

Before we discussing the details of NRLSS method, the noise GMM with K mixtures is trained to improve the model robustness under data-driven framework. Similar to the speech GMM in (1), for a particular noise type denoted by its MFCC feature vector $\mathbf{w}$, we use $m_{\mathbf{w}}$ to denote the index of Gaussian mixture, $N(\mathbf{w}|\mathbf{u}_{m_{\mathbf{w}}}, \mathbf{\Sigma}_{m_{\mathbf{w}}})$ to denote the $m_{\mathbf{w}}^{th}$ Gaussian distribution, $g_{m_{\mathbf{w}}}$ to denote the corresponding mixture weight, where $\mathbf{u}_{m_{\mathbf{w}}}$ and $\mathbf{\Sigma}_{m_{\mathbf{w}}}$ are the mean vector and the covariance matrix, respectively. The noise GMM facilitates the online derivation of noise MGTS $\mathbf{m}_{\mathbf{w}}$.

Now, we describe the proposed NRLSS method. By incorporating noise information to (4), we have

$$\begin{aligned}&\mathbf{m}^t_{\mathbf{x},u:u+\tau_{\max}}, \mathbf{m}^t_{\mathbf{w},v:v+\tau_{\max}} \\ &= \arg\max_{\tau} \max_{\mathbf{m}_{\mathbf{x},u:u+\tau}} p(\mathbf{m}_{\mathbf{x},u:u+\tau}, \mathbf{m}_{\mathbf{w},v:v+\tau}|\mathbf{y}_{t:t+\tau})\end{aligned} \quad (5)$$

where the MFCC feature sequence $\mathbf{y}_{t:t+\tau} = \{\mathbf{y}_n : n = t, t+1, \ldots, t+\tau\}$ represents a noisy speech segment from frame t to frame $t+\tau$. The noise MGTS $\mathbf{m}_{\mathbf{w},v:v+\tau} = \{m_{\mathbf{w},i} : i = v, v+1, \ldots, v+\tau\}$ denotes the noise segment, and $\mathbf{m}^t_{\mathbf{w},v:v+\tau_{\max}}$ denotes the longest matching noise segment (LMNS). Similar to (4), $\mathbf{m}_{\mathbf{x},u:u+\tau}$ and $\mathbf{m}^t_{\mathbf{x},u:u+\tau_{\max}}$ are the clean speech segment and the LMSS at frame t respectively. We note that $p(\mathbf{m}_{\mathbf{x},u:u+\tau}, \mathbf{m}_{\mathbf{w},v:v+\tau}|\mathbf{y}_{t:t+\tau})$ has a good property: the larger it is, the longer $\mathbf{m}_{\mathbf{x},u:u+\tau}$ and $\mathbf{m}_{\mathbf{w},v:v+\tau}$ match, the more specific they become, due to the increase of distinct temporal dynamics, and hence the more accuracy for searching $\mathbf{m}^t_{\mathbf{x},u:u+\tau_{\max}}$ and $\mathbf{m}^t_{\mathbf{w},v:v+\tau_{\max}}$.

For simplicity, we assume that the adjacent frames are conditionally independent and all possible corpus segments have an equal prior probability. Using Bayesian principle, we simplify

$p(\mathbf{m}_{\mathbf{x},u:u+\tau}, \mathbf{m}_{\mathbf{w},v:v+\tau}|\mathbf{y}_{t:t+\tau})$ in (5) as

$$p(\mathbf{m}_{\mathbf{x},u:u+\tau}, \mathbf{m}_{\mathbf{w},v:v+\tau}|\mathbf{y}_{t:t+\tau}) = \frac{p(\mathbf{y}_{t:t+\tau}|\mathbf{m}_{\mathbf{x},u:u+\tau}, \mathbf{m}_{\mathbf{w},v:v+\tau})}{\sum_{u',v'} p(\mathbf{y}_{t:t+\tau}|\mathbf{m}_{\mathbf{x},u':u'+\tau}, \mathbf{m}_{\mathbf{w},v':v'+\tau}) + p(\mathbf{y}_{t:t+\tau}|\phi)} \tag{6}$$

where the denominator of (6) consists of two terms. The first term $\sum_{u',v'} p(\mathbf{y}_{t:t+\tau}|\mathbf{m}_{\mathbf{x},u':u'+\tau}, \mathbf{m}_{\mathbf{w},v':v'+\tau})$ corresponds to a sum of segmental likelihoods over all possible segments of speech and noise corpora, which are likely to match the given $\mathbf{y}_{t:t+\tau}$. The second term $p(\mathbf{y}_{t:t+\tau}|\phi)$ represents a likelihood that the given $\mathbf{y}_{t:t+\tau}$ matches an 'unseen segment'ϕ, which is unlikely constrained in the speech and noise corpora. This likelihood associated with ϕ can be suitably formed on the speech and noise GMMs

$$p(\mathbf{y}_{t:t+\tau}|\phi) = \prod_{n=t}^{\tau} \left[\sum_{m_{\mathbf{x}}=1}^{M} \sum_{m_{\mathbf{w}}=1}^{K} g_{m_{\mathbf{x}}} g_{m_{\mathbf{w}}} N(\mathbf{y}_n|\lambda(m_{\mathbf{x}}), \psi(m_{\mathbf{w}})) \right] \tag{7}$$

where $\lambda(m_{\mathbf{x}}) = \{\mathbf{u}_{m_{\mathbf{x}}}, \mathbf{\Sigma}_{m_{\mathbf{x}}}\}, \psi(m_{\mathbf{w}}) = \{\mathbf{u}_{m_{\mathbf{w}}}, \mathbf{\Sigma}_{m_{\mathbf{w}}}\}$.

In the calculation of $p(\mathbf{y}_{t:t+\tau}|\mathbf{m}_{\mathbf{x},u:u+\tau}, \mathbf{m}_{\mathbf{w},v:v+\tau})$ in (6), the introduction of $\mathbf{m}_{\mathbf{w},v:v+\tau}$ has significantly increased the joint search complexity. To reduce the complexity, given $\mathbf{y}_{t:t+\tau}$ at frame t, we only constrain $\mathbf{m}_{\mathbf{x},u:u+\tau}$, and remain the noise segment denoted by $\mathbf{m}^*_{\mathbf{w},t:t+\tau}$ unconstrained. Therefore, the segmental likelihood can be expressed by

$$p(\mathbf{y}_{t:t+\tau}|\mathbf{m}_{\mathbf{x},u:u+\tau}, \mathbf{m}^*_{\mathbf{w},t:t+\tau}) = \prod_{n=t}^{t+\tau} N(\mathbf{y}_n|\lambda(m_{\mathbf{x},i(n)}), \psi(m^*_{\mathbf{w},n})) \tag{8}$$

where $i(n)$ denotes the most-likely linear warping function between $\mathbf{y}_{t:t+\tau}$ and $\mathbf{m}_{\mathbf{x},u:u+\tau}$. The constrained $\mathbf{m}_{\mathbf{x},u:u+\tau}$ is searched from the speech corpus. And the unconstrained noise segment denoted by $\mathbf{m}^*_{\mathbf{w},t:t+\tau} = \{m^*_{\mathbf{w},t}, m^*_{\mathbf{w},t+1}, \ldots, m^*_{\mathbf{w},t+\tau}\}$ is searched from the noise GMM. More specifically, each term $\mathbf{m}^*_{\mathbf{w},n}, n = t, t+1, \ldots, t+\tau$ is formed by choosing the index of Gaussian mixture to maximize the Gaussian distribution of the given noisy speech. Thus,

$$m^*_{\mathbf{w},n} = \underset{1\le m_{\mathbf{w}}\le K, 1\le m_{\mathbf{x}}\le M}{\arg\max} N(\mathbf{y}_n|\lambda(m_{\mathbf{x}}), \psi(m_{\mathbf{w}})) \tag{9}$$

In our implementations, we use a VTS model to calculate $N(\mathbf{y}_n|\lambda(m_{\mathbf{x}}), \psi(m_{\mathbf{w}}))$ at frame n (To be discussed in Section 3.2).

Above all, to solve the problem of searching $\mathbf{m}^t_{\mathbf{x},u:u+\tau_{\max}}$ and $\mathbf{m}^t_{\mathbf{w},v:v+\tau_{\max}}$ associated with $\mathbf{y}^t_{t:t+\tau_{\max}}$ at frame t, (5) is used, but $\mathbf{m}_{\mathbf{w},v:v+\tau}$ is replaced with $\mathbf{m}^*_{\mathbf{w},t:t+\tau}$.

3.2. Gaussian likelihood calculation with VTS

In (7), (8) and (9), we have to calculate $N(\mathbf{y}|\lambda(m_{\mathbf{x}}), \psi(m_{\mathbf{w}}))$. The noisy speech, clean speech and noise have non-linear relationship in MFCC domain, we cannot directly obtain $N(\mathbf{y}|\lambda(m_{\mathbf{x}}), \psi(m_{\mathbf{w}}))$ given Gaussian distributions of the clean speech and noise. For obtaining this likelihood, we use the well-known VTS [14] approximation algorithm to transform their non-linear relationship into linear.

In this paper, only additive noise is considered and channel distortion is ignored. In the MFCC domain, the noisy speech can be expanded as [14]

$$\mathbf{y} = \mathbf{x} + f(\mathbf{w} - \mathbf{x}) \tag{10}$$

with

$$f(\mathbf{w} - \mathbf{x}) = \mathbf{D}\log[1 + \mathbf{D}^{-1}\exp(\mathbf{w} - \mathbf{x})] \tag{11}$$

where $\mathbf{y}$, $\mathbf{w}$ and $\mathbf{x}$ represent MFCC feature vectors of noisy speech, additive noise, and clean speech respectively. $\mathbf{D}$ denotes the discrete cosine transform (DCT) matrix and $\mathbf{D}^{-1}$ is its pseudo-inverse.

Given a noisy speech, considering the speech and noise mean vectors $(\mathbf{u}_{m_{\mathbf{x}}}, \mathbf{u}_{m_{\mathbf{w}}})$ as the operation points, we can transform the nonlinear formulation in (10) and (11) into linear by using a first-order VTS approximation

$$\mathbf{y} = \mathbf{u}_{m_{\mathbf{x}}} + f(\mathbf{u}_{m_{\mathbf{w}}} - \mathbf{u}_{m_{\mathbf{x}}}) + \mathbf{B}(\mathbf{x} - \mathbf{u}_{m_{\mathbf{x}}}) + \mathbf{G}(\mathbf{w} - \mathbf{u}_{m_{\mathbf{w}}}) \tag{12}$$

with

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}|_{(\mathbf{u}_{m_{\mathbf{x}}}, \mathbf{u}_{m_{\mathbf{w}}})} = \mathbf{B}$$
$$\frac{\partial \mathbf{y}}{\partial \mathbf{w}}|_{(\mathbf{u}_{m_{\mathbf{x}}}, \mathbf{u}_{m_{\mathbf{w}}})} = \mathbf{I} - \mathbf{B} = \mathbf{G} \tag{13}$$

where $\mathbf{B} = \mathbf{D}\mathrm{diag}(\frac{1}{1+\mathbf{D}^{-1}\exp(\mathbf{u}_{m_{\mathbf{x}}} - \mathbf{u}_{m_{\mathbf{w}}})})\mathbf{D}^{-1}$ and $\mathrm{diag}(\cdot)$ denotes the operation for extracting the elements of a column vector to form a diagonal matrix.

Taking the expectations of (12), we can respectively obtain $\mathbf{u}_{\mathbf{y}}$ (the mean of $\mathbf{y}$) and its covariance matrix $\mathbf{\Sigma}_{\mathbf{y}}$ as follows

$$\mathbf{u}_{\mathbf{y}} = \mathbf{u}_{m_{\mathbf{x}}} + f(\mathbf{u}_{m_{\mathbf{w}}} - \mathbf{u}_{m_{\mathbf{x}}}) \tag{14}$$

$$\mathbf{\Sigma}_{\mathbf{y}} = \mathbf{B}\mathbf{\Sigma}_{m_{\mathbf{x}}}\mathbf{B}^T + \mathbf{G}\mathbf{\Sigma}_{m_{\mathbf{w}}}\mathbf{G}^T \tag{15}$$

As a result, the Gaussian distribution of the noisy speech can be calculated by

$$N(\mathbf{y}|\lambda(m_{\mathbf{x}}), \psi(m_{\mathbf{w}}) = N(\mathbf{y}|\mathbf{u}_{\mathbf{y}}, \mathbf{\Sigma}_{\mathbf{y}}) \tag{16}$$

3.3. Spectrum estimation

We present a continuous spectral reconstruction method [16] to form an estimation of the underlying clean speech spectrum $\hat{S}_n(\omega)$ and noise spectrum $\hat{N}_n(\omega)$ at frame n. We describe the method in the following which uses the determined $\mathbf{m}^t_{\mathbf{x},u:u+\tau_{\max}}$ to obtain $\hat{S}_n(\omega)$. The same method can be used to obtain $\hat{N}_n(\omega)$. For the convenient calculation, the noise power spectrum $\hat{N}_n(\omega)^2$ is obtained by Minima Controlled Recursive Averaging (MCRA) algorithm [17]. Both estimated spectra can be used to construct the MWF. So we can estimate $\hat{S}_n(\omega)$ by [16]

$$\hat{S}_n(\omega) = \frac{\sum_t \mathbf{A}(\mathbf{m}^t_{\mathbf{x},i(n)}) p(\mathbf{m}^t_{\mathbf{x},u:u+\tau_{\max}}, \mathbf{m}^*_{\mathbf{w},v:v+\tau_{\max}}|\mathbf{y}_{t:t+\tau_{\max}})}{\sum_t p(\mathbf{m}^t_{\mathbf{x},u:u+\tau_{\max}}, \mathbf{m}^*_{\mathbf{w},v:v+\tau_{\max}}|\mathbf{y}_{t:t+\tau_{\max}})} \tag{17}$$

where $\mathbf{m}^t_{\mathbf{x},i(n)}$ is a speech corpus frame associated with $\mathbf{y}_{t:t+\tau_{\max}}$ from $\mathbf{m}^t_{\mathbf{x},u:u+\tau_{\max}}$ calculated by (5), and $\mathbf{A}(\mathbf{m}^t_{\mathbf{x},i(n)})$ obtained by (3) denotes the pre-recorded magnitude spectrum of clean speech corresponding to the $\mathbf{m}^t_{\mathbf{x},i(n)}$. The frames within the same segment share a common weight $p(\mathbf{m}^t_{\mathbf{x},u:u+\tau_{\max}}, \mathbf{m}^*_{\mathbf{w},t:t+\tau_{\max}}|\mathbf{y}_{t:t+\tau_{\max}})$, which is obtained by (5). The denominator of (17) is a normalization term. By taking an average among all possible matching segments, we can achieve a smooth estimation over successive LMSS. This improved the accuracy of $\hat{S}_n(\omega)$ in current frame for the inaccurate segment matching.

3.4. Modified wiener filter

Conventional WF for speech enhancement could improve the quality of speech to a certain extent, however, there still remains a high-level noise under silence or unvoiced segments. To solve that problem, we introduce a MWF and its transfer function can be written as

$$H_n(\omega) = \frac{(1-\rho)\hat{S}_n(\omega)^2}{(1-\rho)\hat{S}_n(\omega)^2 + \rho\hat{N}_n(\omega)^2} \tag{18}$$

where ρ is NCCC between noisy and noise spectra in [15].

In order to offer a good balance between the speech estimation accuracy for voiced segments and noise reduction for the silence or unvoiced segments, we employ the MWF only when the value of NCCC is larger than an empirical value and the conventional WF is used otherwise.

4. Performance evaluation

Performance of the proposed speech enhancement method is evaluated in this section. The LRTDs modeling of speech is trained with one hour speech database, i.e., a speech GMM, with M=512 mixtures, whose value is a trade-off between enhancement performance and computation complexity. The dimension of MFCC feature is set to 42. And the frame length is 20ms with a frame shift of 10ms. The test speech is chosen from NTT database including 8 sentences from 4 female speakers and 4 male speakers. The length of each sentence sampled at 8 kHz is 8s. In our experiments, noise signals are selected from Noisex-92 including white noise, street noise, office noise and babble noise. The GMMs for the first three noise signals, the mixture number of GMM is 8, and for the babble noise, mixture of GMM is 16. The input SNR is defined as 0dB, 5dB and 10dB, respectively.

To evaluate the proposed speech enhancement method from different perspectives, three objective evaluation measures, i.e. the average segmental signal-to-noise ratio (SSNR) [18], average log-spectral distortion (LSD) [19], and perceptual evaluation of speech quality (PESQ) measures [20], are employed. The performance of the proposed method (PM) is evaluated and compared with three reference methods, including WF [1] method, WEDM [4] and CB [5] methods. In CB method [5], the speech and noise codebook sizes are 1024 and 8 respectively, except for babble noise with codebook size equals to 16. Table 1, 2, and 3 show the results of PESQ, LSD and SSNR, respectively. And the score of the best performing algorithm is shown in bold-face letters.

Table 1: *PESQ of Respective Enhancement Algorithms.*

Noise type	Input SNR	Method				
		Noisy	WF	WEDM	CB	PM
white	0dB	1.39	2.19	2.02	2.20	**2.34**
	5dB	1.61	2.47	2.36	2.46	**2.70**
	10dB	2.01	2.71	2.64	2.65	**3.02**
babble	0dB	1.72	1.93	2.00	1.86	**2.08**
	5dB	2.04	2.34	2.40	2.26	**2.60**
	10dB	2.44	2.70	2.75	2.55	**2.99**
office	0dB	2.00	2.22	2.29	2.32	**2.42**
	5dB	2.40	2.60	2.65	2.66	**2.89**
	10dB	2.75	2.94	2.81	2.95	**3.20**
street	0dB	2.31	2.80	**2.84**	2.79	2.72
	5dB	2.66	3.04	3.07	3.05	**3.18**
	10dB	2.95	3.26	3.29	3.31	**3.43**

Table 2: *LSD Results.*

Noise type	Input SNR	Method				
		Noisy	WF	WEDM	CB	PM
white	0dB	19.76	10.15	11.65	9.8	**9.23**
	5dB	17.51	8.79	10.13	8.35	**5.77**
	10dB	15.33	7.43	8.72	6.96	**5.16**
babble	0dB	15.67	11.53	11.59	10.70	**9.23**
	5dB	13.60	9.66	9.75	9.17	**7.95**
	10dB	9.38	7.88	7.98	7.69	**6.32**
office	0dB	13.69	9.16	9.48	9.15	**7.05**
	5dB	11.71	7.46	7.78	7.61	**5.70**
	10dB	9.85	5.90	6.15	6.17	**5.08**
street	0dB	13.19	7.77	8.21	7.85	**7.04**
	5dB	11.25	6.12	6.57	6.38	**5.67**
	10dB	9.46	**4.67**	5.06	5.06	5.14

Table 3: *SSNR Improvement Results.*

Noise type	Input SNR	Method				
		Noisy	WF	WEDM	CB	PM
white	0dB	-	17.08	14.50	12.41	**22.15**
	5dB	-	15.85	13.37	11.63	**18.80**
	10dB	-	14.50	12.16	10.68	**16.28**
babble	0dB	-	7.57	7.61	10.51	**13.93**
	5dB	-	7.10	7.14	9.19	**12.26**
	10dB	-	6.49	6.53	7.80	**11.08**
office	0dB	-	11.18	10.67	13.37	**19.99**
	5dB	-	10.80	10.28	12.07	**19.44**
	10dB	-	10.34	9.82	10.66	**18.36**
street	0dB	-	14.22	13.44	18.05	**23.00**
	5dB	-	12.25	11.47	14.76	**21.59**
	10dB	-	11.71	10.98	13.51	**19.71**

Table 1 shows the PESQ scores for the input noisy speech and for the reconstructed speech from the enhancement methods. Since the reference methods do not consider the temporal dynamics in a long-range period, the PM performed more natural output and better speech quality than all these methods.

Similar conclusions can be drawn based on the results comparisons of LSD and SSNR. Highly significant improvements indicate that our method achieve better speech quality while suppress fluctuant background noise much more effective than reference methods. In table 2, an exception that the LSD of PM is higher than the references for street noise at 10 dB is caused by the inaccurate segment searching for few frames.

The improvement of the proposed method come at a cost of an increased complexity, which dominated by the high-order GMM training.

5. Conclusions

In this paper, a speech enhancement method based upon data-driven speech segmentation and recognition scheme is proposed. The PM has an advantage that the enhanced speech has the potential to be almost free of artifacts due to the pre-stored 'clean'signal from speech corpus. It improves speech quality by exploiting the NRLSS with VTS approximation algorithm. Moreover, the introduction of the MWF helps further reduce the residue background noise. The objective results show that our method has better performance than reference methods.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No.61471014).

7. References

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 857–869, 2005.

[5] S. Srinivasan, J. Samuelsson, and W. B. Klejin, "Codebook-based bayesian speech enhancement for nonstationary environments," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 441–452, 2007.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788 – 791, 1999.

[7] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 882–892, 2007.

[8] Z. Z. Gao, C. C. Bao, F. Bao, and M. S. Jia, "HMM-based speech enhancement using vector taylor series and parallel modeling in mel-frequency domain," in *Signal Processing, Communications and Computing (ICSPCC), 2014 IEEE International Conference on*, 2014, pp. 733–737.

[9] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 6, no. 5, pp. 445–455, 1998.

[10] X. Xiao and R. M. Nickel, "Speech enhancement with inventory style speech resynthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1243–1257, 2010.

[11] J. Ming, "Maximizing the continuity in segmentation - a new approach to model, segment and recognize speech," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 3849–3852.

[12] J. Ming, R. Srinivasan, and D. Crookes, "A corpus-based approach to speech enhancement from nonstationary noise," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 822–836, 2011.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society Series B (statistical Methodology)*, vol. 39, no. 1, pp. 1–38, 1977.

[14] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector taylor series for noisy speech recognition," in *Proc. Int. Conf. on Spoken Language Processing*, 2009, pp. 229–232.

[15] F. Bao, H. J. Dou, M. S. Jia, and C. C. Bao, "Speech enhancement based on a few shapes of speech spectrum," in *Signal and Information Processing (ChinaSIP), 2014 IEEE China Summit International Conference on*, 2014, pp. 90–94.

[16] J. Ming, R. Srinivasan, D. Crookes, and A. Jafari, "CLOSE-a data-driven approach to speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 7, pp. 1355–1368, 2013.

[17] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *Signal Processing Letters, IEEE*, vol. 9, no. 1, pp. 12–15, 2002.

[18] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. New York: Prentice Hall, West Nyack, New York, U.S.A, 1988.

[19] A. Abramson and I. Cohen, "Simultaneous detection and estimation approach for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2348–2359, 2007.

[20] ITU-T, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs ITU-T Rec. P.862*, 2001.