A Gain-adaptive Parallel HMM for Speech Enhancement

Qi He and Chang-chun Bao
Speech and Audio Signal Processing Laboratory, School of Electronic Information and Control Engineering,
Beijing University of Technology, Beijing, China, 100124
very1990@emails.bjut.edu.cn, baochch@bjut.edu.cn

***Abstract*—This document is an example of what your final camera-ready manuscript to APSIPA ASC 2015 should look like. Authors are asked to conform to the directions reported in this document. The key problem in Hidden Markov model (HMM)-based speech enhancement is how to obtain an appropriate weighted Wiener filter (WWF). This paper presents a gain-adaptive parallel HMM (PHMM) for speech enhancement based on Mel-frequency spectral (MFS) features and linear prediction (LP) coefficients of speech and noise, and a WWF modified by the speech-presence probability (SPP) is developed. MFS-HMM (i.e., the HMM in MFS domain) and LP-HMM (i.e., the HMM in LP domain) constitute the proposed PHMM, which is obtained by the parallel training method. The forward probabilities in all mixtures of states for noisy speech MFS-HMM are calculated as the weighting factors of WWF. In order to obtain more accurate noisy speech MFS-HMM and solve the mismatching problem of spectral energy between the training and test signals in MFS domain, we introduce two gain factors to adaptively adjust the spectral energy of speech and noise, respectively. The gain factors are estimated online by the expectation maximization (EM) algorithm. The pre-trained LP-HMMs of speech and noise only contain the spectral shapes of speech and noise, so the EM algorithm is also employed to estimate LP spectral gains for constructing Wiener filters in the proposed WWF. The evaluation results confirm the superiority of the proposed method.**

I. INTRODUCTION

Speech enhancement aims at suppressing noise while improving the quality and intelligibility of speech. Speech enhancement algorithm has been widely used in many speech signal processing applications, such as mobile communication, hearing aids and speech recognition systems.

Research works on single-channel speech enhancement have been conducted over decades and various methods have been proposed. Generally, conventional methods can be roughly divided into three categories: spectral subtraction method [1][2], Wiener filtering method [3][4], and statistical-model-based method [5][6]. These methods share the common drawback that the de-noising performance is well in stationary noise environments but not suitable for non-stationary noises. The main reason is that they cannot track the quick variation of non-stationary noise energy.

To solve the problem aforementioned, some methods using the priori information about speech and noise have been developed, such as autoregressive (AR) HMM-based speech enhancement methods [7][8][9]. For these methods, speech and noise are modeled as separate AR processes for a given state. The AR-HMMs of speech and noise were trained offline using recorded signals. And a WWF was employed to enhance noisy speech. The weighting factors of Wiener filters in WWF were calculated using the AR-HMM of noisy speech, which can be estimated based on the pre-trained AR-HMMs of speech and noise. Each Weiner filter corresponding to the relative weighting factor was constructed by the estimated LP envelope spectra of speech and noise. The AR-HMM-based speech enhancement methods could track the change of noise energy and remove amount of noise, but neglecting the residual noise between the harmonics of estimated speech, typically due to the inaccuracy fitting of spectra between the harmonics of voiced segments in estimated speech. Some residual noises were still remained in the enhanced speech. T H. Veisi et al thought that MFS feature is superior for the estimation of weighting factors [10]. In order to improve the performance of de-noising and the accuracy of estimation for weighting factors, researcher proposed a HMM-based speech enhancement method in MFS and spectral amplitude domains [10][11]. For these works, the HMMs of speech and noise in MFS domain were trained for calculating the weighting factors of Wiener filters more precisely. Simultaneously, the additional HMMs of speech and noise in spectral amplitude domain were obtained by the parallel training method [10] for constructing the Wiener filters. Although these kinds of methods could obtain more accurate weighting factors and suitable spectral fitting, they do not explicitly consider the mismatching problem of spectral energy between the training and test signals, which will result in the inaccurate estimation of speech and noise spectra and weaker robustness.

Besides, conventional HMM-based methods ignore the sparseness property of speech signal in the time-frequency domain. The study has shown that the SPP often equals to zero in some frequency bins or time frames only containing noise, such as the frequency bins between the harmonics of noisy speech. If the HMM-based Wiener filter is modified appropriately by the SPP, the noise in these frequency bins or in those frames could be removed efficiently.

In this paper, we propose a unified solution to the aforesaid problems using a parallel MFS and LP coefficients HMM. Firstly, the parallel training method is used to obtain the proposed PHMMs of speech and noise. Each PHMM consists of the MFS-HMM and the LP-HMM. Secondly, two

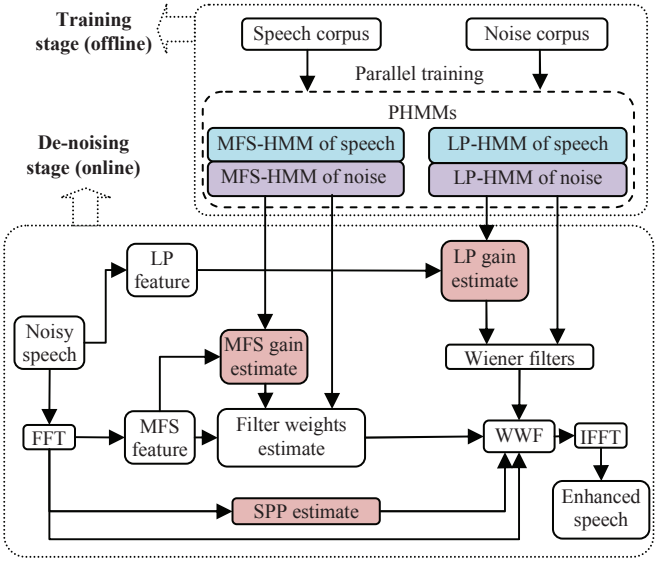

Fig.1.The diagram of proposed speech enhancement system

gain adjustment factors are estimated online using the EM algorithm [12] to solve the mismatching problem of spectral energy in MFS domain. Thirdly, we obtain the noisy MFS-HMM using the estimated gain factors and the pre-trained MFS-HMMs of speech and noise. Furthermore, the forward algorithm [11] is applied to calculate the weighting factors of WWF based on noisy MFS-HMM. The LP-HMMs of speech and noise only contain the spectral shapes of speech and noise, so the LP spectral gains are estimated by the EM algorithm online. Each Wiener filter of WWF is constructed by the spectral shapes and corresponding LP gains of speech and noise. At last, the WWF combined with the SPP is developed to further remove residual noise in estimated speech.

The proposed speech enhancement system is shown in Fig.1. The system consists of two stages: one is the offline training stage which obtains the PHMMs of both speech and noise using the parallel training method; another is the real-time de-noising stage, using the developed WWF modified by the SPP. In the system, the symbols, FFT and IFFT stand for Fast Fourier transforming and inverse FFT, respectively.

The remainder of this paper is organized as follows: In Sec.2, we briefly describe the training method of PHMM. In Sec.3, the proposed speech enhancement method is introduced in details. Our experiments and results are shown in Sec.4. Finally, Sec.5 concludes the paper.

II. THE TRAINING METHOD OF PHMM

In this section, the training method is applied to the PHMMs of both speech and noise.

The process of MFS feature extraction is shown in Fig.2.

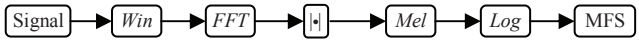

Fig.2.The block diagram of MFS feature extraction

where *Win*, $|\bullet|$, *Mel* and *Log* denote the operations of windowing, magnitude spectrum calculation, Mel-filtering, and logarithm, respectively.

Let superscript notations *mfs* and *lpc* correspond to the MFS and LP domains, respectively. We define the parameters set of MFS-HMM as follows:

$$\boldsymbol{\theta}^{mfs}=\{N,M,\boldsymbol{\pi}^{mfs},\mathbf{a}^{mfs},\mathbf{c}^{mfs},\boldsymbol{\mu}^{mfs},\boldsymbol{\Sigma}^{mfs}\} \quad (1)$$

where N and M are the number of states and mixtures, respectively, $\boldsymbol{\pi}^{mfs}=\{\pi_i\}$ is the set of initial state probabilities for state i, $\mathbf{a}^{mfs}=\{a_{ij}\}$ is the set of state transition probabilities from state i to state j, $\mathbf{c}^{mfs}=\{\mathrm{c}_{m|i}\}$ is the set of mixture weighting value in mixture m for state i, $\boldsymbol{\mu}^{mfs}=\{\boldsymbol{\mu}_{m|i}^{mfs}\}$ and $\boldsymbol{\Sigma}^{mfs}=\{\boldsymbol{\Sigma}_{m|i}^{mfs}\}$ are the set of mean vectors $\boldsymbol{\mu}_{m|i}^{mfs}$ and diagonal covariance matrices $\boldsymbol{\Sigma}_{m|i}^{mfs}$ of Gaussian density function in mixture m for state i, respectively.

The parameters set of LP-HMM can be defined as $\boldsymbol{\theta}^{lpc}=\{\boldsymbol{\mu}^{lpc}\}$ where $\boldsymbol{\mu}^{lpc}=\{\boldsymbol{\mu}_{m|i}^{lpc}\}$ is set of LP coefficients mean vectors $\boldsymbol{\mu}_{m|i}^{lpc}$ in mixture m for state i. Note that the LP-HMM and MFS-HMM share the common states and mixtures in the proposed PHMM.

Fig.3 shows the parallel training scheme of PHMM. Firstly, we get the MFS features and LP coefficients of speech or noise from training corpuses frame by frame. Then in MFS domain, the standard Baum-Welch algorithm [13] is used to train the MFS-HMM until model converges.

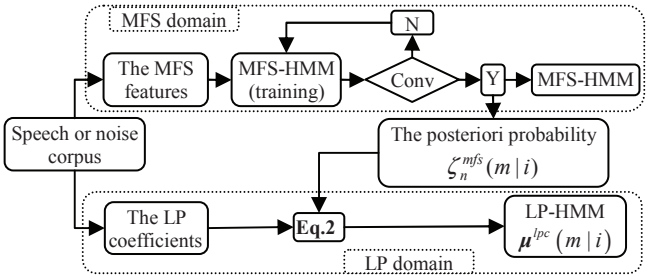

Fig.3.The training diagram of PHMM

Following that, the mean vector of LP-HMM $\boldsymbol{\mu}_{m|i}^{lpc}$ in LP domain can be obtained directly by:

$$\boldsymbol{\mu}_{m|i}^{lpc}=\frac{\sum_{n=1}^{F}\zeta_n^{mfs}(m\,|\,i)\mathbf{v}_n^{lpc}}{\sum_{n=1}^{F}\zeta_n^{mfs}(m\,|\,i)} \quad (2)$$

where $\zeta_n^{mfs}(m|i)$ is the posteriori probability in mixture m for state i from MFS-HMM in nth frame, $\mathbf{v}_n^{lpc}=[\alpha_n^0,\ldots,\alpha_n^q]^T$ is the vector of LP coefficients with q being the LP-model order in nth frame, and F is the total number of training frames.

III. HELPFUL HINTS

As known, the general structure of HMM-based WWF $W_{sum}(k)$ can be written as:

$$W_{sum}(k)=\sum_s p(s\,|\,\mathbf{y}_{0:n})W_s(k)\,;\qquad 0\le k\le K-1 \quad (3)$$

with

$$W_s(k)=\frac{P_{x,s}(k)}{P_{x,s}(k)+P_{w,s}(k)} \tag{4}$$

where K is the FFT size, $s=(\bar{s},\ddot{s})$ is the noisy state-mixture, which consists of the speech state-mixture $\bar{s}=(i_x,m_x)$ and the noise state-mixture $\ddot{s}=(i_w,m_w)$, $W_s(k)$ is the Wiener filter in state-mixture s, which is constructed by $P_{x,s}(k)$ and $P_{w,s}(k)$ corresponding to the estimated spectra of speech and noise in state-mixture s, respectively, $\mathbf{y}_{0:n}$ is the observed noisy speech from the 0th frame to nth frame, the filter weighting factor $p(s|\mathbf{y}_{0:n})$, i.e., the conditional probability of noisy state-mixture s can be obtained using the forward algorithm, the ranges for speech state i_x, speech mixture m_x, noise state i_w, and noise mixture m_w are defined as $1\le i_x\le N_x$, $1\le m_x\le M_x$, $1\le i_w\le N_w$ and $1\le m_w\le M_w$, respectively.

Obviously, a better WWF can be achieved by providing sufficiently accurate weighting factor $p(s|\mathbf{y}_{0:n})$ and corresponding Wiener filter $W_s(k)$.

A. Noisy Speech Model

In this paper, the noisy speech $\mathbf{y}_n$ is composed of clean speech $\mathbf{x}_n$ and additive noise signal $\mathbf{w}_n$ in nth frame. And let $\tilde{\mathbf{x}}_n$ and $\tilde{\mathbf{w}}_n$ denote the clean speech and noise signal from training corpuses, respectively.

Considering the mismatching problem of spectral energy, we employ the following noisy speech model where speech and noise are assumed to be independent:

$$\mathbf{y}_n=\mathbf{x}_n+\mathbf{w}_n=g_{x_n}\tilde{\mathbf{x}}_n+g_{w_n}\tilde{\mathbf{w}}_n \tag{5}$$

where g_{x_n} and g_{w_n} are the introduced gain adjustment factors in nth frame. For simplicity, we drop the label n in following part.

According to Fig.2 and Eq.5, we can obtain the following equation in MFS domain:

$$\mathbf{y}^{mfs}(l)=\log\left(\mathbf{y}^{mel}(l)\right)=\log\left(g_x\tilde{\mathbf{x}}^{mel}(l)+g_w\tilde{\mathbf{w}}^{mel}(l)\right);\quad 1\le l\le L \tag{6}$$

where superscript *mel* denotes the feature after Mel-filtering, L denotes the order of Mel-filtering.

The PHMM of noisy speech can be estimated based on the pre-trained PHMMs of speech and noise. Each noisy composite state i_y consists of combination of the speech state i_x and the noise state i_w, i.e. $i_y=(i_x, i_w)$. This is analogous for the noisy mixture m_y, i.e. $m_y=(m_x, m_w)$. And we have:

$$\begin{aligned}
&N_y=N_xN_w \quad , \quad M_y=M_xM_w\\
&\pi_{i_y}=\pi_{i_x}\pi_{i_w};\\
&a_{i_yi'_y}=a_{i_xi'_x}a_{i_wi'_w};\quad 1\le i'_x\le N_x,\ 1\le i'_w\le N_w\\
&\mathrm{c}_{m_y|i_y}=\mathrm{c}_{m_x|i_x}\,\mathrm{c}_{m_w|i_w};
\end{aligned} \tag{7}$$

where N_y and M_y denote the number of states and mixtures in noisy speech PHMM, respectively.

For each noisy state-mixture $s=(\bar{s},\ddot{s})=(i_x,m_x,i_w,m_w)$, using the first order vector Taylor series expansion around the point $\{\tilde{\boldsymbol{\mu}}^{mel}_{x,\bar{s}}(l),\tilde{\boldsymbol{\mu}}^{mel}_{w,\ddot{s}}(l)\}$ for Eq.5, we can obtain that:

$$\begin{aligned}
\mathbf{y}^{mfs}_s(l)=&\log\left(g_{x,s}\tilde{\boldsymbol{\mu}}^{mel}_{x,\bar{s}}(l)+g_{w,s}\tilde{\boldsymbol{\mu}}^{mel}_{w,\ddot{s}}(l)\right)\\
&+\frac{g_{x,s}\left(\tilde{\mathbf{x}}^{mel}_{\bar{s}}(l)-\tilde{\boldsymbol{\mu}}^{mel}_{x,\bar{s}}(l)\right)+g_{w,s}\left(\tilde{\mathbf{w}}^{mel}_{\ddot{s}}(l)-\tilde{\boldsymbol{\mu}}^{mel}_{w,\ddot{s}}(l)\right)}{g_{x,s}\tilde{\boldsymbol{\mu}}^{mel}_{x,\bar{s}}(l)+g_{w,s}\tilde{\boldsymbol{\mu}}^{mel}_{w,\ddot{s}}(l)}
\end{aligned} \tag{8}$$

By taking the expectation and variance on both side of the Eq.8, the lth element of the mean vector and the lth diagonal element of diagonal covariance matrix in the noisy MFS-HMM can be shown as:

$$\begin{aligned}
\boldsymbol{\mu}^{mfs}_{y,s}(l)&=\log\left(g_{x,s}\tilde{\boldsymbol{\mu}}^{mel}_{x,\bar{s}}(l)+g_{w,s}\tilde{\boldsymbol{\mu}}^{mel}_{w,\ddot{s}}(l)\right)\\
\boldsymbol{\Sigma}^{mfs}_{y,s}(l)&=\frac{g^2_{x,s}\tilde{\boldsymbol{\Sigma}}^{mel}_{x,\bar{s}}(l)+g^2_{w,s}\tilde{\boldsymbol{\Sigma}}^{mel}_{w,\ddot{s}}(l)}{\left(g_{x,s}\tilde{\boldsymbol{\mu}}^{mel}_{x,\bar{s}}(l)+g_{w,s}\tilde{\boldsymbol{\mu}}^{mel}_{w,\ddot{s}}(l)\right)^2}
\end{aligned} \tag{9}$$

with [11]

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}^{mel}_{x,\bar{s}}(l)&=\exp\left(\tilde{\boldsymbol{\mu}}^{mfs}_{x,\bar{s}}(l)+0.5\tilde{\boldsymbol{\Sigma}}^{mfs}_{x,\bar{s}}(l)\right)\\
\tilde{\boldsymbol{\Sigma}}^{mel}_{x,\bar{s}}(l)&=\tilde{\boldsymbol{\mu}}^{mel}_{x,\bar{s}}(l)\tilde{\boldsymbol{\mu}}^{mel}_{x,\bar{s}}(l)\left[\exp\left(\tilde{\boldsymbol{\Sigma}}^{mfs}_{x,\bar{s}}(l)\right)-1\right]\\
\tilde{\boldsymbol{\mu}}^{mel}_{w,\ddot{s}}(l)&=\exp\left(\tilde{\boldsymbol{\mu}}^{mfs}_{w,\ddot{s}}(l)+0.5\tilde{\boldsymbol{\Sigma}}^{mfs}_{w,\ddot{s}}(l)\right)\\
\tilde{\boldsymbol{\Sigma}}^{mel}_{w,\ddot{s}}(l)&=\tilde{\boldsymbol{\mu}}^{mel}_{w,\ddot{s}}(l)\tilde{\boldsymbol{\mu}}^{mel}_{w,\ddot{s}}(l)\left[\exp\left(\tilde{\boldsymbol{\Sigma}}^{mfs}_{w,\ddot{s}}(l)\right)-1\right]
\end{aligned} \tag{10}$$

where $\tilde{\boldsymbol{\mu}}^{mfs}_{x,\bar{s}}(l)$, $\tilde{\boldsymbol{\Sigma}}^{mfs}_{x,\bar{s}}(l)$ and $\tilde{\boldsymbol{\mu}}^{mfs}_{w,\ddot{s}}(l)$, $\tilde{\boldsymbol{\Sigma}}^{mfs}_{w,\ddot{s}}(l)$ come from the pre-trained MFS-HMMs of PHMMs of speech and noise, respectively.

B. Gain Factors Estimation in MFS Domain

In order to obtain the mean vector and diagonal covariance matrix of noisy MFS-HMM, the gain adjustment factors in Eq.9 should be determined online. Here the EM algorithm [12] is applied to estimate $g_{x,s}$ and $g_{w,s}$. For simplicity, we drop the state-mixture label s.

For the speech component, the maximization step in the EM algorithm is to find a new g_x that maximize the auxiliary function $L(g_x\,|\,\hat{g}_x^{j-1})$ from the expectation step, as shown in Eq.11, where j denotes the iteration index.

$$L(g_x\,|\,g_x^{j-1})=\int f(\mathbf{x}^{mfs}_n\,|\,\mathbf{y}^{mfs}_n,g_x{}^{j-1})\log f(\mathbf{x}^{mfs}_n\,|\,g_x)\,\mathrm{d}\,\mathbf{x}^{mfs}_n \tag{11}$$

with

$$f(\mathbf{x}^{mfs}_n\,|\,g_x)=\frac{1}{(2\pi)^{L/2}\sqrt{|\boldsymbol{\Sigma}^{mfs}_x|}}\exp\left\{-\frac{(\mathbf{x}^{mfs}_n-\boldsymbol{\mu}^{mfs}_x)^T(\boldsymbol{\Sigma}^{mfs}_x)^{-1}(\mathbf{x}^{mfs}_n-\boldsymbol{\mu}^{mfs}_x)}{2}\right\} \tag{12}$$

where T denotes the Hermitian transpose. Considering

$$\mathbf{X}^{mfs}=\log(g_x\tilde{\mathbf{X}}^{mel})=(\log g_x)\boldsymbol{e}+\tilde{\mathbf{X}}^{mfs} \tag{13}$$

with $\boldsymbol{e}=[1,\,1,\,\ldots,1]^T_{1\times L}$. By taking the expectation and variance on both side of the Eq.13, we can get

$$\begin{aligned}
\boldsymbol{\mu}^{mfs}_x&=\log(g_x)\boldsymbol{e}+\tilde{\boldsymbol{\mu}}^{mfs}_x\\
\boldsymbol{\Sigma}^{mfs}_x&=\tilde{\boldsymbol{\Sigma}}^{mfs}_x
\end{aligned} \tag{14}$$

Let $g'_x=\log(g_x)$. By substituting Eq.14 into Eq.12, we have:

$$\begin{aligned}
&\log f(\mathbf{x}^{mfs}_n\,|\,g_x)\\
&=-\frac{L}{2}\log(2\pi)-\frac{1}{2}\log\left|\tilde{\boldsymbol{\Sigma}}^{mfs}_x\right|\\
&\quad-\frac{(\mathbf{x}^{mfs}_n-g'_x\boldsymbol{e}-\tilde{\boldsymbol{\mu}}^{mfs}_x)^T(\tilde{\boldsymbol{\Sigma}}^{mfs}_x)^{-1}(\mathbf{x}^{mfs}_n-g'_x\boldsymbol{e}-\tilde{\boldsymbol{\mu}}^{mfs}_x)}{2}
\end{aligned} \tag{15}$$

Differentiating Eq.11 with g'_x and setting the resulting expression to zero, we can obtain the update equation of g'_x for the *j*th iteration as follows:

$$\boldsymbol{e}^T(\tilde{\boldsymbol{\Sigma}}_x^{mfs})^{-1}\int f(\mathbf{x}_n^{mfs}\mid \boldsymbol{y}_n^{mfs}, g_x{}^{j-1})(\mathbf{x}_n^{mfs}-g_x'^{j}\boldsymbol{e}-\tilde{\boldsymbol{\mu}}_x^{mfs})d\mathbf{x}_n^{mfs}=0 \quad (16)$$

By simplifying the above formula, we then get

$$\boldsymbol{e}^T(\tilde{\boldsymbol{\Sigma}}_x^{mfs})^{-1}(\tilde{\boldsymbol{R}}_x^{mfs}-g_x'^{j}\boldsymbol{e}-\tilde{\boldsymbol{\mu}}_x^{mfs})=0 \quad (17)$$

with

$$\begin{aligned}\tilde{\mathbf{R}}_x^{mfs}&=E(\mathbf{x}_n^{mfs}\mid \mathbf{y}_n^{mfs},\hat{g}_x{}^{j-1})\\&=\int f(\mathbf{x}_n^{mfs}\mid \mathbf{y}_n^{mfs},\hat{g}_x{}^{j-1})\mathbf{x}_n^{mfs}d\mathbf{x}_n^{mfs}\\&=g_x'^{j-1}\boldsymbol{e}+\tilde{\boldsymbol{\mu}}_x^{mfs}+\frac{\tilde{\boldsymbol{\Sigma}}_x^{mfs}\left(\boldsymbol{\Sigma}_y^{mfs}\right)^{-1}\left(\mathbf{y}_n^{mfs}-\boldsymbol{\mu}_y^{mfs}\right)}{1+\exp\left(g_w'^{j-1}\boldsymbol{e}+\tilde{\boldsymbol{\mu}}_w^{mfs}-g_x'^{j-1}\boldsymbol{e}-\tilde{\boldsymbol{\mu}}_x^{mfs}\right)}\end{aligned} \quad (18)$$

which actually denotes the Bayesian minimum mean square error estimate of $\mathbf{x}_n^{mfs}$ [14]. The update equation of g'^{j}_w has the similar structure as Eq.17 with *x* replaced by *w*.

Getting the log-gain factors $g'_{x,s}$ and $g'_{w,s}$ for each state-mixture *s*, we can obtain the mean vectors and diagonal covariance matrices of noisy MFS-HMM by substituting $g_{x,s}=\exp(g'_{x,s})$ and $g_{w,s}=\exp(g'_{w,s})$ into the Eq.9. Then the filter weighting factor $p(s|\mathbf{y}_{0:n})$ in Eq.3 is calculated by using the forward algorithm.

C. LP Gains Estimation

In the proposed WWF, the Wiener filter $W_s(k)$ in each noisy state-mixture *s* is constructed by the LP spectra of speech and noise. It is written as:

$$W_s(k)=\frac{\dfrac{g_{x,s}^{lpc}}{\left|A_{x,\bar{s}}(k)\right|^2}}{\dfrac{g_{x,s}^{lpc}}{\left|A_{x,\bar{s}}(k)\right|^2}+\dfrac{g_{w,s}^{lpc}}{\left|A_{w,\ddot{s}}(k)\right|^2}} \quad (19)$$

where $1/\left|A_{x,\bar{s}}(k)\right|^2$ and $1/\left|A_{w,\ddot{s}}(k)\right|^2$ denote the spectral shapes of speech and noise in the speech state-mixture $\bar{s}$ and the noise state-mixture $\ddot{s}$,respectively, $g_{x,s}^{lpc}$ and $g_{w,s}^{lpc}$ are the respective LP spectral gains of speech and noise in the state-mixture *s*. The spectral shapes of speech and noise are stored in the LP-HMMs of speech and noise, respectively, which is obtained by the parallel training method offline. So we need to estimate the LP spectral gains online. The EM algorithm is also applied to estimate the LP spectral gains in each state-mixture *s*.

Omit the label *s*. Let $\mathbf{z}_n$ denotes the hidden speech or noise signal in *n*th frame. Let g_z^{lpc} denotes the random variable corresponding to the LP spectral gain of speech or noise. The auxiliary function can be shown as:

$$C(g_z^{lpc}\mid g_z^{lpc,j-1})=\int f(\mathbf{z}_n\mid \mathbf{y}_n,g_z^{lpc,j-1})\log[f(\mathbf{z}_n\mid g_z^{lpc})]\mathrm{d}\mathbf{z}_n \quad (20)$$

where $f(\mathbf{z}_n\mid g_z^{lpc})$ is defined as follows [9]:

$$f(\mathbf{z}_n\mid g_z^{lpc})=\frac{1}{(2\pi g_z^{lpc})^{K/2}\sqrt{|\mathbf{D}_z|}}\exp(-\frac{1}{2g_z^{lpc}}\mathbf{z}_n^T(\mathbf{D}_z)^{-1}\mathbf{z}_n) \quad (21)$$

where the covariance matrix $\mathbf{D}_z=(\mathbf{A}_z^T\mathbf{A}_z)^{-1}$, where $\mathbf{A}_z$ is a *K*K* lower triangular Toeplitz matrix with the LP coefficients $[1,\alpha_{z_1},...,\alpha_{z_p},0,...,0]^T$ as the first column. Then we have:

$$\begin{aligned}&\log f(\mathbf{z}_n\mid g_z^{lpc})\\&=-\frac{K}{2}\log\left(2\pi g_z^{lpc}\right)-\frac{1}{2}\log(|\mathbf{D}_z|)-\frac{1}{2g_z^{lpc}}\mathbf{z}_n^T(\mathbf{D}_z)^{-1}\mathbf{z}_n\end{aligned} \quad (22)$$

By differentiating Eq.20 with g_z^{lpc} and setting the result to zero, we can obtain:

$$\int f(\mathbf{z}_n\mid \mathbf{y}_n,g_z^{lpc,j-1})\left(\frac{\mathbf{z}_n^T(\mathbf{D}_z)^{-1}\mathbf{z}_n}{2(g_z^{lpc})^2}-\frac{K}{2g_z^{lpc}}\right)\mathrm{d}\mathbf{z}_n=0 \quad (23)$$

Multiplying by $(g_z^{lpc})^2$ on both side of Eq.23 and simplifying the result, we get:

$$g_z^{lpc}=\frac{\int f(\mathbf{z}_n\mid \mathbf{y}_n,g_z^{lpc,j-1})\mathbf{z}_n^T(\boldsymbol{D}_z)^{-1}\mathbf{z}_n\,\mathrm{d}\mathbf{z}_n}{K} \quad (24)$$

According to the Eq.24, the updated equations of speech and noise LP spectral gains for the *j*th iteration can be given by:

$$g_x^{lpc,j}=\frac{V_x^{j-1}}{K}\ ;\ g_w^{lpc,j}=\frac{V_w^{j-1}}{K} \quad (25)$$

with

$$\begin{aligned}V_x^{j-1}&=\int f(\mathbf{x}_n\mid \mathbf{y}_n,g_x^{lpc,j-1})\mathbf{x}_n^T(\mathbf{D}_x)^{-1}\mathbf{x}_n d\mathbf{x}_n\\V_w^{j-1}&=\int f(\mathbf{w}_n\mid \mathbf{y}_n,g_w^{lpc,j-1})\mathbf{w}_n^T(\mathbf{D}_w)^{-1}\mathbf{w}_n d\mathbf{w}_n\end{aligned} \quad (26)$$

which can be calculated by referring to the method in [9].

D. The Structure of Developed WWF

The conventional AR-HMM-based speech enhancement method cannot remove the noise between the harmonics of noisy speech. Therefore, we introduce the SPP to modify the Wiener filter of WWF in each state-mixture *s* for suppressing the noise.

Let $H_1(k)$ and $H_0(k)$ denote the event that speech is present and absent in frequency bin *k*, respectively. Assuming the complex Gaussian distribution of speech and noise FFT coefficients [5], we can get the conditional PDFs of the observed signal in noisy state-mixture *s* by:

$$\begin{aligned}p_s(Y(k)\mid H_0(k))&=\frac{1}{\pi\lambda_{w,s}(k)}\exp\left(-\frac{|Y(k)|^2}{\lambda_{w,s}(k)}\right)\\p_s(Y(k)\mid H_1(k))&=\frac{1}{\pi\left(\lambda_{w,s}(k)+\lambda_{x,s}(k)\right)}\exp\left(-\frac{|Y(k)|^2}{\lambda_{w,s}(k)+\lambda_{x,s}(k)}\right)\end{aligned} \quad (27)$$

where $Y(k)$ is the FFT coefficient of noisy speech, $\lambda_{x,s}(k)=E_s\left\{|X(k)|^2\right\}$ and $\lambda_{w,s}(k)=E_s\left\{|W(k)|^2\right\}$ represent the variances of speech and noise in noisy state-mixture *s,* respectively.

Using Bayes rule, we can express the posteriori SPP in noisy state-mixture *s* as follows:

$$\begin{aligned}&p_s\left(H_1(k)\mid Y(k)\right)\\&=\frac{p_s(Y(k)\mid H_1(k))p(H_1(k))}{p_s(Y(k)\mid H_0(k))p(H_0(k))+p_s(Y(k)\mid H_1(k))p(H_1(k))}\end{aligned} \quad (28)$$

By substituting Eq.27 into the above formula and letting $\beta(k) = p(H_1(k)) = 1 - p(H_0(k))$, the posteriori SPP can be obtained by [15]:

$$p_s(H_1(k)\,|\,Y(k)) = \frac{\beta(k)}{\beta(k) + (1-\beta(k))(1+\xi'_s(k))\exp(-\nu'_s(k))} \quad (29)$$

with

$$\xi'_s(k) = \frac{g^{lpc}_{x,s} \Big/ \left|A_{x,\bar{s}}(k)\right|^2}{P_{w,mcra}(k)\beta(k)} \quad , \quad \nu'_s(\mathrm{k}) = \frac{\xi'_s(k)P_y(k)}{(1+\xi'_s(k))P_{w,mcra}(k)} \quad (30)$$

where $g^{lpc}_{x,s}$ is the estimate of speech LP gain in noisy state-mixture s, $P_y(k)$ is the observed noisy speech spectrum and the noise spectrum $P_{w,mcra}(k)$ is obtained using the Minima Controlled Recursive Averaging (MCRA) algorithm [16].

We utilize a monotonically increasing sigmoid function to describe the range of $\beta(k)$ as follows [17]:

$$\beta(k) = \frac{1}{\left(1+\exp\left(3.22 - \frac{P_y(k)}{P_{w,mcra}(k)}\right)\right)} \quad (31)$$

In order to further attenuate noise in noise-only frame, we get a probability coefficient ρ which can be formulated as:

$$\rho = \frac{1}{N}\sum_{k=0}^{N-1} p_{mcra}(H_1(k)) \quad (32)$$

where $p_{mcra}(H_1(k))$ is a smoothed SPP for frequency bin k in MCRA algorithm .

Fig.4 depicts the variance of coefficient ρ for the clean speech which is degraded by white noise in the SNR level of 10dB. From the Fig.4, we can observe that the probability coefficient ρ approximates to 0 and 1 in silence and voiced segments, respectively. The ρ is not equal to 1 because the voiced segments contain noise. And for those frames only containing noise, the ρ approximates to 0.

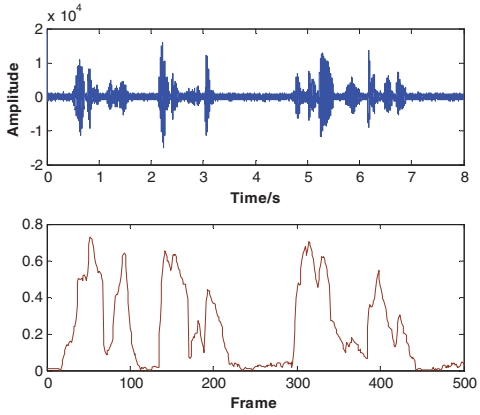

Fig.4. the variation of ρ *vs.* noisy speech (white noise, SNR=10dB)

By applying the posteriori SPP and the probability coefficient ρ, we get the Wiener filter combined with SPP in noisy state-mixture s as follows:

$$W'_s(k) = \frac{p_s(H_1(k)\,|\,X(k)) \cdot \rho \cdot \dfrac{g^{lpc}_{x,s}}{\left|A_{\bar{s}}(k)\right|^2}}{\rho \cdot \dfrac{g^{lpc}_{x,s}}{\left|A_{\bar{s}}(k)\right|^2} + (1-\rho) \cdot \dfrac{g^{lpc}_{w,s}}{\left|A_{\bar{s}}(k)\right|^2}} \quad (33)$$

where $g^{lpc}_{w,s}$ is noise LP gain in state-mixture s, $1/|A_{\bar{s}}(k)|^2$ is the spectral shape of the mean vector $\boldsymbol{\mu}^{lpc}_{\bar{s}}$ from noise LPC-HMM in state-mixture $\bar{s}$. For ρ>0.5, i.e. speech component is more than noise component in current frame, the $W'_s(k)$ can reserve more speech component than $W'_s(k)$ without ρ. For ρ<0.5, the $W'_s(k)$ can remove more noise. The final modified WWF is

$$W'_{sum}(k) = \sum_s p(s\,|\,\mathbf{y}_{0:n}) W'_s(k) \quad (34)$$

IV. EXPERIMENTS AND EVALUATIONS

This section discusses the performance evaluation of the proposed method. The test clean speech is extracted from NTT database and resampled to 8*kHz*. The length of each utterance is 8s. Four types of noise are selected from Noisex-92 database, including white, babble, factory, and f16 noises. The clean speech test set is degraded by adding these noise types in three input SNR levels, which is defined as 0dB, 5dB and 10dB. The test materials contain 20 utterances from 10 female speakers and 10 male speakers. The frame length is 32ms (256 samples) and overlapped for 16ms (128 samples). The samples are windowed using normalized Hamming window. The FFT size is 256. Tenth-order LP analysis is adopted for both speech and noise. The dimension of MFS vector is 24. The speech PHMM has 8 states and 16 mixture components per state. The length of clean speech for training is 20 minutes. Each noise PHMM has 4 states and 4 mixture components per state. And the length of each noise for training is 4 minutes.

The objective measurements used in the evaluations are segmental SNR (SSNR)[18], log-spectral distortion (LSD)[19], and the perceptual evaluation of speech quality (PESQ)[20]. The reference methods for objective evaluations are the AR-HMM-based method (Ref. A) [9] and the MFS-HMM-based method (Ref. B) [10].

A. The SSNR evaluation.

The SSNR is often applied to evaluate the de-noising performance of speech enhancement method. It is computed as:

$$SSNR = \frac{1}{T'}\sum_{t=1}^{T'} 10\log_{10}\left(\frac{\sum_{n=1}^{F'} x^2(t,n)}{\sum_{n=1}^{F'}\left(x(t,n)-\hat{x}(t,n)\right)^2}\right) \quad (35)$$

where t is the frame index, T' is the total number of frames in test utterance, F' is the length of frame, $x(t,n)$ denotes the clean speech, $\hat{x}(t,n)$ denotes the enhanced speech or noisy speech.

Table.1 shows the SSNR improvements of various methods. We can observe that the de-noising performance of our method is better than reference methods in majority testing conditions. The Ref. B can get the higher SNR improvement in 10dB input SNR, but its performance is unstable and declines sharply with the decreasing of input SNR because of the problem of energy mismatching. And the

proposed method achieves higher average SSNR improvement than reference methods.

TABLE.1. TEST RESULTS OF SSNR IMPROVEMENT

Noise	SNR	Noisy	Ref. A	Ref. B	Proposed
White	0dB	−−	15.86	3.80	20.01
	5dB	−−	15.00	10.55	18.56
	10dB	−−	13.68	17.51	16.84
Babble	0dB	−−	11.87	5.59	14.56
	5dB	−−	10.18	12.46	13.61
	10dB	−−	8.47	14.88	12.95
Factory	0dB	−−	14.67	13.57	18.15
	5dB	−−	13.07	15.45	17.26
	10dB	−−	11.42	16.10	15.91
F16	0dB	−−	14.67	6.96	18.74
	5dB	−−	12.88	11.24	17.98
	10dB	−−	11.38	17.36	16.70
Average	**0dB**	−−	**14.27**	**7.48**	**17.87**
	5dB	−−	**12.78**	**12.43**	**16.85**
	10dB	−−	**11.24**	**16.46**	**15.60**

B. The LSD evaluation

The LSD between clean speech and enhanced speech is given by:

$$LSD = \frac{1}{T'}\sum_{t=1}^{T'}\sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left(10\log_{10}\left(\frac{X(t,k)}{\hat{X}(t,k)}\right)\right)^2} \tag{36}$$

where $X(t,k)$ is the power spectrum of clean speech, $\hat{X}(t,k)$ is the power spectrum of enhanced speech or noisy speech.

Table.2 gives the LSD test results. The average LSD of the proposed method is lower than reference methods, which demonstrate that the proposed method causes lower speech component distortion than reference methods.

TABLE.2. TEST RESULTS OF LSD

Noise	SNR	Noisy	Ref. A	Ref. B	Proposed
White	0dB	18.52	10.46	16.18	7.21
	5dB	16.40	8.97	10.69	5.89
	10dB	14.38	7.63	6.69	4.99
Babble	0dB	15.34	10.28	12.06	8.76
	5dB	13.42	8.84	7.32	7.26
	10dB	11.64	7.45	6.15	5.97
Factory	0dB	14.95	9.37	13.61	6.67
	5dB	13.03	7.86	8.98	5.43
	10dB	11.24	6.46	8.26	4.67
F16	0dB	16.69	10.73	8.48	7.09
	5dB	14.66	9.20	6.85	5.47
	10dB	12.77	7.52	6.3	4.66
Average	**0dB**	**16.38**	**10.21**	**12.58**	**7.43**
	5dB	**14.38**	**8.72**	**8.46**	**6.01**
	10dB	**12.51**	**7.27**	**6.85**	**5.07**

C. The PESQ evaluation

The PESQ reflects the perceptual quality of enhanced speech. The higher PESQ value indicates the better perceptual quality of speech. Experimental results for PESQ are shown in Table.3.

TABLE.3. TEST RESULTS OF PESQ

Noise	SNR	Noisy	Ref. A	Ref. B	Proposed
White	0dB	1.47	1.89	1.58	2.20
	5dB	1.73	2.19	2.38	2.66
	10dB	2.11	2.43	2.92	2.99
Babble	0dB	1.71	1.91	1.97	1.98
	5dB	2.02	2.29	2.41	2.42
	10dB	2.41	2.60	2.76	2.78
Factory	0dB	1.79	2.08	2.30	2.38
	5dB	2.18	2.41	2.59	2.78
	10dB	2.59	2.67	2.86	3.09
F16	0dB	1.58	1.80	1.89	2.15
	5dB	1.89	2.24	2.52	2.70
	10dB	2.33	2.52	2.91	3.07
Average	**0dB**	**1.64**	**1.92**	**1.94**	**2.18**
	5dB	**1.96**	**2.28**	**2.48**	**2.64**
	10dB	**2.36**	**2.56**	**2.86**	**2.98**

From the Table.3, we can find that the proposed method leads to higher average PESQ value than reference methods in all input SNRs, which means that the proposed method can obtain better auditory quality.

D. The spectrogram comparison

The Fig.5 and Fig.6 show the spectrograms of enhanced speech obtained by various speech enhancement methods. The Fig.5 corresponds to the clean test speech which is degraded by the white noise in SNR level of 10dB. And the Fig.6 corresponds to the SNR level of 5dB.

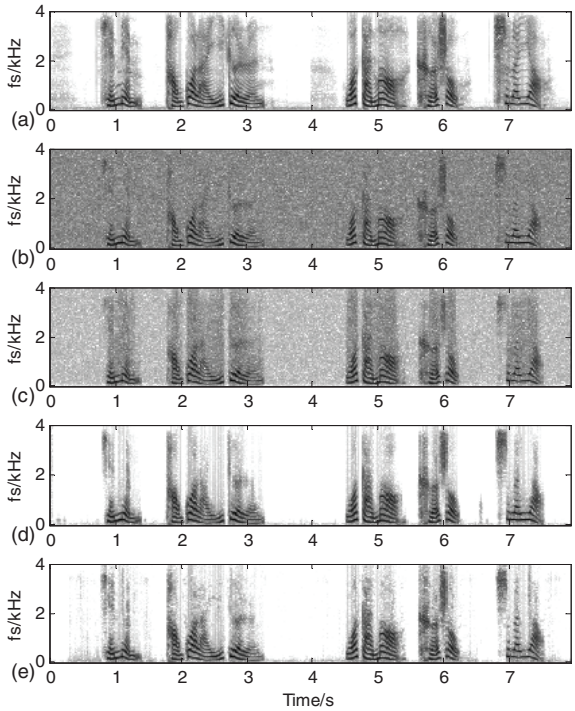

Fig.5. Spectrograms of (a) clean speech, (b) noisy speech (white noise, input SNR=10dB), (c) Ref. A, (d) Ref. B, (e) the proposed method.

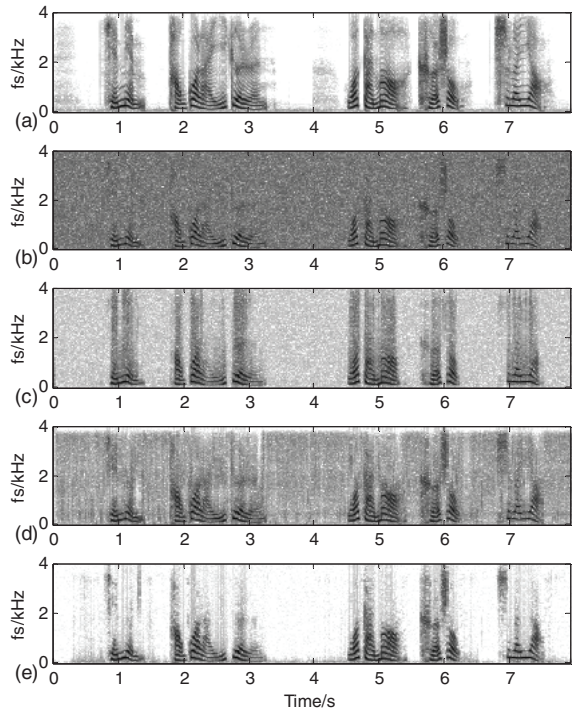

Fig.6. Spectrograms of (a) clean speech, (b) noisy speech (white noise, input SNR=5dB), (c) Ref. A, (d) Ref. B, (e) the proposed method

The spectrograms shown in Fig.5 and Fig.6 demonstrate the effectiveness of the proposed method in time-frequency perspective. From the Fig.5 to Fig.6, we can observe that all speech enhancement methods can suppress the background noise to some extent. The proposed method performs better than reference methods. For the noisy speech with input SNR level of 10dB, the proposed and Ref. B methods have the similar de-noising performance and can remove more noise than Ref. A method. Nevertheless, for the noisy speech with input SNR level of 5dB, the de-noising performance of Ref. B method reduces sharply due to the problem of energy mismatching. In comparison with the Ref. B method, the proposed method has the better robustness. In addition, Fig.5 (e) and Fig.6 (e) illustrate the residual noise between harmonics of noisy speech can be eliminated partly by the proposed method. In general, the proposed method outperforms conventional methods in nearly all (different SNRs and noise environments) conditions.

V. CONCLUSIONS

A gain-adaptive PHMM is investigated for speech enhancement based on MFS feature and LP coefficient. The proposed PHMM is used successfully to speech and noise modeling in this paper. Two gain factors in MFS domain are introduced to solve the mismatching problem of spectral energy between training and test signals. The EM algorithm is utilized to obtain the online gain factors. And we apply the SPP to combine with the HMM-constrained WWF to achieve the goal of removing much more background noise without speech component distortion. The evaluations show the superiority of the proposed method in comparison with the references.

ACKNOWLEDGMENT

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

REFERENCES

[1] J. S. Lim, A. V. Oppenheim, ”Enhancement and bandwidth compression of noisy speech,” *Proc. IEEE*, Vol 67, pp.1586-1604, Dec. 1979.

[2] S. Boll. “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-27, pp. 113-120.Apr. 1979.

[3] R. J. McAulay and K. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[4] S. V. Vaseghi, Advanced Signal Processing and Digital Noise Reduction. New York: Wiley, 1998, ch.6.

[5] Y. Ephraim and D. Malah, “Speech Enhancement Using a Minimum Mean Square Error Short-time Spectral Amplitude Estimator”, *IEEE Tran. Acoust., Speech Signal Process*, 32(6), 1109-1121.1984.

[6] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech Signal Process*. , vol. 33, pp. 443-445, 1985.

[7] Y. Ephraim, “A Bayesian estimation approach for speech enhancement using hidden Markov models,” *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725-735, Apr. 1992.

[8] Sameti, H., Sheikhzadeh, H., Deng, L. & Brennan, R.L., “HMM-Based Strategies for Enhancement of Speech Signals Embedded in Nonstationary Noise”, *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 5, pp. 445, 1998.

[9] David Y. Zhao, and W. B. Kleijn, HMM-based gain modeling for enhancement of speech in noise, *IEEE Trans. Audio, Speech, Lang. Process*, vol. 15, no. 3, pp. 163–176, Jan.2006.

[10] H. Veisi and H. Sameti, “A Parallel Cepstral and Spectral Modeling for HMM-based Speech Enhancement,” *Digital Signal Processing (DSP), 2011 17th International Conference on. IEEE*, 2011, pp.1-6.

[11] Zheng-zheng Gao, Chang-chun Bao, Feng Bao, and Mao-shen Jia. HMM-Based Speech Enhancement Using Vector Taylor Series and Parallel Modeling in Mel-Frequency Domain. *Signal Processing, Communications & Computing* (*ICSPCC, 2014*), Guilin, Guangxi, China, Aug 5-8, 2014.

[12] J. A. Bilmes, “A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models,” *Univ. of Berkeley, Berkeley, CA, Tech. Rep. ICSI-TR-97-021*, 1997.

[13] Rabiner, L.R., “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257-286, 1989.

[14] Jinyu Li, Seltzer, M.L. and Yifan Gong, “Improvement to VTS feature enhancement,” in *Proc. ICASSP*, 2012, pp.4677-4680.

[15] Cohen I., “Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator,” *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 112–116, Apr. 2002.

[16] Cohen, I. “Noise estimation by minima controlled recursive averaging for robust speech enhancement,” *IEEE Signal Process. Lett.*, 9(1), 12-15. 2002.

[17] Zhong-hua Fu and Jhing-Fa Wang, "Speech presence probability estimation based on integrated time-frequency minimum tracking for speech enhancement in adverse environments", *Acoustics Speech and Signal Processing* (*ICASSP*). *IEEE*, Mar. 2010, pp. 4258-4261.
[18] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, "Objective measures of speech quality," Englewood Cliffs, NJ: Prentice Hall, 1988.
[19] Abramson, A., Cohen, I., "Simultaneous Detection and Estimation Approach for Speech Enhancement", *IEEE Trans. Speech Audio Process*. , 15(8), pp. 2348-2359, 2007.
[20] "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," *ITU-T Recommendation*, P.862, Feb, 2001.