Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 205–220

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Speech enhancement using hidden Markov models in Mel-frequency domain

Hadi Veisi *, Hossein Sameti

Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

Received 3 January 2012; received in revised form 31 July 2012; accepted 6 August 2012
Available online 24 August 2012

Abstract

Hidden Markov model (HMM)-based minimum mean square error speech enhancement method in Mel-frequency domain is focused on and a parallel cepstral and spectral (PCS) modeling is proposed. Both Mel-frequency spectral (MFS) and Mel-frequency cepstral (MFC) features are studied and experimented for speech enhancement. To estimate clean speech waveform from a noisy signal, an inversion from the Mel-frequency domain to the spectral domain is required which introduces distortion artifacts in the spectrum estimation and the filtering. To reduce the corrupting effects of the inversion, the PCS modeling is proposed. This method performs concurrent modeling in both cepstral and magnitude spectral domains. In addition to the spectrum estimator, magnitude spectrum, log-magnitude spectrum and power spectrum estimators are also studied and evaluated in the HMM-based speech enhancement framework.

The performances of the proposed methods are evaluated in the presence of five noise types with different SNR levels and the results are compared with several established speech enhancement methods especially auto-regressive HMM-based speech enhancement. The experimental results for both subjective and objective tests confirm the superiority of the proposed methods in the Mel-frequency domain over the reference methods, particularly for non-stationary noises.

Keywords: HMM-based speech enhancement; Mel-frequency; Parallel cepstral and spectral (PCS)

1. Introduction

Speech signals need to be enhanced in many applications for various purposes such as boosting overall speech quality, increasing intelligibility or improving the performance of speech coding and speech recognition systems. Although speech enhancement has been studied in various aspects, single microphone speech enhancement is of wide interest due to the extent and variety of its applications. Furthermore, it has remained as a challenging topic over the years and numerous researchers have tried to suggest solutions for this problem. Among various approaches proposed to estimate clean speech from a noisy signal, statistical speech enhancement has shown a promising perspective and has received a great deal of attention during the past two decades (Ephraim and Malah, 1984, 1985; Porter and Boll, 1984; Martin, 2005; Chen and Loizou, 2007; Ephraim et al., 1989; Ephraim, 1992a–c; Sameti, 1994; Sameti et al., 1998; Logan, 1998; Zhao and Kleijn, 2007; Srinivasan et al., 2007). Statistical speech enhancement methods can be roughly categorized into two classes, *non-model-based* and *model-based* methods. In most of speech applications, speech signal is assumed a stochastic process that is stationary within a small segment of signal called a frame. In the non-model-based methods, certain statistical models are assumed a priori for speech and noise signals within a frame. In this approach, noise statistics and clean speech are estimated online using an estimator such as the minimum mean square error (MMSE). Speech and noise parameters are frequently assumed to be Gaussian distributed (Ephraim and Malah, 1984, 1985); although recent estimators assume Laplacian or Gamma distribution for

* Corresponding author. Fax: +98 2166551525.
E-mail addresses: veisi@ce.sharif.edu (H. Veisi), sameti@sharif.edu (H. Sameti).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.08.005

speech parameters (Porter and Boll, 1984; Martin, 2005; Chen and Loizou, 2007). Well-known examples of the non-model-based methods are Wiener filter (Lim and Oppenheim, 1978), short-time spectral amplitude (STSA) estimator (Ephraim and Malah, 1984), log-spectral amplitude (LSA) estimator (Ephraim and Malah, 1985) and the MMSE estimator using non-Gaussian priors (Porter and Boll, 1984; Martin, 2005; Chen and Loizou, 2007).

In the model-based approach, the modeling is done using the statistics of the signal over multiple frames. This modeling is performed using hidden Markov model (HMM) Ephraim et al., 1989; Ephraim, 1992a–c; Sameti, 1994; Sameti et al., 1998; Logan, 1998; Zhao and Kleijn, 2007, Gaussian mixture model (GMM), or codebook-based methods (Srinivasan et al., 2007). HMM-based speech enhancement is the renowned model-based technique and resolves the common problems of classical speech enhancement methods such as the spectral subtraction method in dealing with rapid variations of the noise characteristics (Sameti et al., 1998) and in generating "musical" noise (Sameti, 1994). Speech enhancement using HMM was originally proposed by Ephraim et al. (1989) based on auto-regressive Gaussian states HMM (AR-HMM) and was extended in (Ephraim, 1992b,c; Sameti, 1994; Sameti et al., 1998; Logan, 1998; Zhao and Kleijn, 2007). In AR-HMM speech enhancement framework, speech and noise are modeled as separate AR processes within a given HMM state and the AR parameters (i.e., the linear predictive coding, LPC, coefficients) are applied as the prior information of the signals. In the Ephraim's works (Ephraim et al., 1989; Ephraim, 1992a,b), the formulation for the clean speech estimation was given in the HMM framework using the MMSE and MAP (maximum a posteriori) estimators. In (Ephraim et al., 1989; Ephraim, 1992a,b), the noise statistics were modeled using a single Gaussian probability density function (pdf). The framework was extended in (Sameti et al., 1998) to allow mixture components in the HMM for noise signals in order to handle non-stationary noises. The method of Sameti et al. (1998) further employed several noise models for different noise environments and the appropriate noise model was chosen using a heuristic noise selection method. In (Zhao and Kleijn, 2007), the problem of energy mismatch between the statistics of the trained models and the observations was studied. This problem had been investigated in other studies (Ephraim, 1992a,b; Sameti et al., 1998), but an explicit HMM-based modeling for both speech and noise gains was proposed in (Zhao and Kleijn, 2007).

In this paper, HMM-based speech enhancement using MMSE estimator is studied. As shown in (Ephraim, 1992c), under certain assumptions, the MMSE estimator results in a weighted sum filtering of the noisy signal. Therefore, accurate estimations of the filter values and filter weights are the main challenges. The filter values are estimated using the parameters of the noise and speech HMMs and the filter weights are computed based on the likelihood of noisy speech signal to the noisy speech model. Consequently, the estimation of the filter weights i.e., the filter selection, can be viewed as a pattern recognition problem in which a higher recognition rate results in more accurate filter selection. Accordingly, in contrast with the established methods which assume auto-regressive process within HMM states, i.e., AR-HMM (Ephraim, 1992b,c; Sameti, 1994; Sameti et al., 1998; Logan, 1998; Zhao and Kleijn, 2007), we study Mel-frequency cepstral coefficients for signal representation in the HMM-based framework to select the filters accurately. This approach is motivated by the prevalent and successful use of cepstral coefficients in automatic speech recognition (ASR) applications due to their effective uncorrelated signal representation. Furthermore, it is known from Ephraim (1992a) and Zhao and Kleijn (2007) that the AR-HMM does not model the spectral fine structure of the voiced speech signal and results in low-level rumbling noise in certain voiced segments. In this study, the speech and noise signals are represented by Mel-frequency spectral (MFS) or Mel-frequency cepstral (MFC) features. MFC denotes the Mel-frequency cepstral coefficients (MFCC) that are normalized using cepstral mean subtraction (CMS). Normalizing the MFCC features using CMS improves the robustness of these features in representing noisy signals. To estimate noisy model parameters from the speech and noise models, PC-PMC (Veisi and Sameti, 2011) method is used which requires feature conversion from the Mel-frequency to the spectral domain. This conversion is a lossy process and results in distortion of the spectrum and inaccurate estimation of filter values in the speech enhancement framework. To overcome this problem, a parallel cepstral and spectral (PCS) modeling is proposed in which the parameters of the signal are concurrently modeled in the magnitude spectral and cepstral domains. In this paper, the HMM-based speech enhancement framework is further extended to employ magnitude, log-magnitude and power spectrum estimators that are optimal in the MMSE sense. This extension is motivated by the fact that the estimation of magnitude spectrum (Ephraim and Malah, 1984), log-magnitude spectrum (Ephraim and Malah, 1985) and power spectrum (Wolfe and Godsill, 2003) were shown to be more effective compared to the spectrum estimation.

The approach conducted in this paper is in line with the model-based feature enhancement methods used for making ASR systems robust to noise (Yu et al., 2008; Segura et al., 2001; Stouten et al., 2006; Sasou et al., 2006; Arakawa et al., 2006); however, our approach is different in the intention and application. By applying this approach, a new application for the Mel-frequency domain HMM is introduced where the improvement of the perceptual quality of a noisy signal is its goal while the focus of the robustness techniques is on the reduction of word error rate of the ASR systems.

The proceeding parts of this paper are organized as follows. In Section 2, the HMM-based speech enhancement in the Mel-frequency domain is explained and the construction of the noisy model is presented. The estimation of

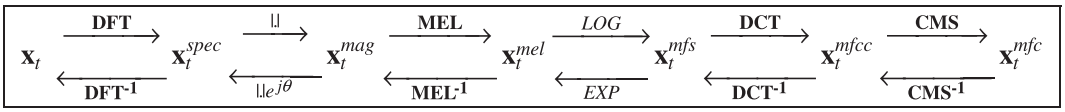


Fig. 1. Operations needed to transform a time-domain vector to the CMS-normalized Mel-frequency cepstral domain and vice versa.

the clean speech employing different estimators is also described in this section. In Section 3, the problem of the inversion of the Mel-frequency domain parameters to the spectral domain is investigated and the PCS modeling is proposed. Section 4 presents the experimental evaluations, and finally, the conclusions are drawn in Section 5.

2. Formulation of HMM-based speech enhancement in Mel-frequency domain

Let $\mathbf{s}_t \in \Re^L$ denotes an L-dimensional time-domain windowed frame of clean speech signal, $\mathbf{d}_t \in \Re^L$ indicates an L-dimensional time-domain uncorrelated additive noise signal, and $\mathbf{y}_t = \mathbf{s}_t + \mathbf{d}_t$ refers to the consequent observed noisy speech frame. Fig. 1 shows sequences of operations needed to transform the time-domain observation vector $\mathbf{x}_t$ (where $x \in \{s, d, y\}$) to the MFC domain (the CMS normalized MFCC) and vice versa. The sequential operations are framing and windowing, discrete Fourier transform (DFT), magnitude calculation, Mel-scaled filter bank (MEL) computation, logarithm (LOG), discrete cosine transform (DCT) and cepstral mean subtraction (CMS). As shown, the signal frames obtained after applying the DFT, MEL, LOG, DCT and CMS are referred to as $\mathbf{x}_t^{spec} \in \Re^{L_{spec}}$, $\mathbf{x}_t^{mel} \in \Re^{L_{mel}}$, $\mathbf{x}_t^{mfs} \in \Re^{L_{mfs}}$, $\mathbf{x}_t^{mfcc} \in \Re^{L_{mfcc}}$ and $\mathbf{x}_t^{mfc} \in \Re^{L_{mfc}}$, respectively. In these notations, L_{spec}, L_{mel}, L_{mfs}, L_{mfcc} and L_{mfc} denote the length of the frames in the specified domains.

To build up the HMM-based speech enhancement system in the Mel-frequency domain, two HMMs are to be trained for clean speech and for noise in the Mel-frequency domain. A clean speech HMM is defined as $\lambda = \{\boldsymbol{\pi}, \mathbf{a}, \mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, where $\boldsymbol{\pi} = \{\pi_i = p[q_1 = \alpha_i]\}$; $1 \leqslant i \leqslant M$ is the set of initial state probabilities, $\mathbf{a} = \{a_{ij} = p[q_{t+1} = \alpha_j | q_t = \alpha_i]\}$; $1 \leqslant i, j \leqslant M$ is the set of state transition probabilities, $\mathbf{c} = \{c_{ki} = p[u_t = \beta_k | q_t = \alpha_i]\}$; $1 \leqslant i \leqslant M$, $1 \leqslant k \leqslant N$ is the set of mixture weights corresponding to the state and mixture pair$\{\alpha_i, \beta_k\}$, and $\{\boldsymbol{\mu}_{\beta_k|\alpha_i}, \boldsymbol{\Sigma}_{\beta_k|\alpha_i}\}$ is the set of Gaussian density parameters, i.e., mean vector and covariance matrix. In these notations, $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_M\}$ and $\beta = \{\beta_1, \beta_2, \ldots, \beta_N\}$ denote the individual states and mixtures, respectively. The variables q_t and u_t indicate the state and mixture, respectively, at time t. M and N denote the number of states and mixtures, respectively. In this model, the transitions between states are assumed to have first order Markovian property. The clean speech HMM in the MFC domain is denoted as λ^{mfc}.

Similarly to the clean speech HMM, a noise HMM having $\tilde{N} \times \tilde{M}$ pdfs, denoted as $\tilde{\lambda}^{mfc}$, is trained for the environmental noise in the MFC domain ($\tilde{M}$ is the number of noise states and $\tilde{N}$ indicates the number of mixture components per state). Therefore, the procedures for the modeling of speech and noise signals are identical except in the number of pdfs. To estimate the parameters of λ^{mfc} and $\tilde{\lambda}^{mfc}$, standard maximum likelihood estimation is performed using Baum re-estimation method (Rabiner, 1989). Unlike the left-to-right HMMs used in speech recognition tasks, for the HMM-based speech enhancement purpose, the global characteristics of the speech are accommodated in a single ergodic HMM with no constraint on the temporal order of the HMM states. Furthermore, in contrast with the HMMs for ASR application in which numbers of HMMs are trained for different acoustic units (such as phonemes or tri-phones), only a single HMM is trained for speech in the HMM-based speech enhancement. Therefore, the enhancement system requires less training data than the ASR system. This is also true for the modeling of noise signal.

For the MFS features, similar procedures with similar assumptions are conducted. In this case, the HMM parameters of the clean speech and noise are referred to as λ^{mfs} and $\tilde{\lambda}^{mfs}$, respectively.

2.1. Constructing noisy speech HMM

The HMM parameters of noisy speech in the MFC domain are estimated as $\bar{\lambda}^{mfc} = \lambda^{mfc} \oplus \tilde{\lambda}^{mfc}$ where $\oplus$ is a combination operator that approximates the noisy model parameters from the speech and noise statistics. The combination operator $\oplus$ depends on various factors such as the manner that noise distorts speech, the features used to represent the noise and speech characteristics, and the pdf assumed for the speech and noise models. Based on our assumptions, i.e., using MFC features, independence and additivity of noise, and Gaussian pdfs for both speech and noise, the combination operator $\oplus$ is a modified version of parallel model combination called PC-PMC (Veisi and Sameti, 2011). PC-PMC utilizes the principal component analysis (PCA) and CMS in parallel model combination (PMC) framework (Gales, 1995). However, in this paper only the CMS normalization (without PCA transform) is applied to the MFCC features. To estimate $\bar{\lambda}^{mfc}$ using PC-PMC, the parameters of speech and noise are primarily transformed from the MFC domain to the MFS (i.e., log-spectral) domain using Eq. (1), and then to the MEL domain utilizing Eq. (2). In these equations, superscripts indicate the domains, $\mathbf{C}^{-1}$ is the inverse (or pseudo inverse) of discrete cosine transform, $(\cdot)^{tr}$ denotes the Hermitian transpose, and $\mathbf{m}^{mfcc} \in \Re^{L_{mfcc}}$ is an approximation of CMS vector in the cepstral domain. This approximation is

a constant vector for all data and is calculated by averaging over all frames within the training set. The CMS vectors can be taken as a random variable and its simplification to a constant vector affect the performance of this robustness technique (Veisi and Sameti, 2011). $\boldsymbol{\mu}(i)$ indicates the ith component of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}(i,j)$ denotes the covariance of the ith and jth components in the feature vector. The Eqs. (1) and (2) are applied on both clean speech and noise model parameters for every mixture component. The combination of speech and noise statistics which results in noisy parameters is given by Eq. (3), where $\{\boldsymbol{\mu}^{mel}, \boldsymbol{\Sigma}^{mel}\}$, $\{\tilde{\boldsymbol{\mu}}^{mel}, \tilde{\boldsymbol{\Sigma}}^{mel}\}$ and $\{\bar{\boldsymbol{\mu}}^{mel}, \bar{\boldsymbol{\Sigma}}^{mel}\}$ indicate the sets of clean speech, noise and noisy mean and covariance parameters in the MEL domain, respectively.

$$\boldsymbol{\mu}^{mfs} = \mathbf{C}^{-1} \cdot (\boldsymbol{\mu}^{mfc} + \mathbf{m}^{mfcc}); \quad \boldsymbol{\Sigma}^{mfs} = \mathbf{C}^{-1} \cdot \boldsymbol{\Sigma}^{mfc} \cdot (\mathbf{C}^{-1})^{tr} \tag{1}$$

$$\begin{aligned} &\boldsymbol{\mu}^{mel}(i) = \exp(\boldsymbol{\mu}^{mfs}(i) + \boldsymbol{\Sigma}^{mfs}(i,i)/2); \\ &\boldsymbol{\Sigma}^{mel}(i,j) = \boldsymbol{\mu}^{mel}(i) \cdot \boldsymbol{\mu}^{mel}(j) \cdot [\exp(\boldsymbol{\Sigma}^{mfs}(i,j)) - 1]; \\ &\qquad 1 \leqslant i,j \leqslant L_{mfs} = L_{mel} \end{aligned} \tag{2}$$

$$\bar{\boldsymbol{\mu}}^{mel} = \boldsymbol{\mu}^{mel} + \tilde{\boldsymbol{\mu}}^{mel}; \bar{\boldsymbol{\Sigma}}^{mel} = \boldsymbol{\Sigma}^{mel} + \tilde{\boldsymbol{\Sigma}}^{mel} \tag{3}$$

The noisy speech parameters $\{\bar{\boldsymbol{\mu}}^{mel}, \bar{\boldsymbol{\Sigma}}^{mel}\}$, are then transformed back to the MFS and MFC domains using the inversion of Eqs. (2) and (1), respectively. This transformation is performed using Eqs. (4) and (5). In Eq. (5) $\bar{\mathbf{m}}^{mfcc}$ denotes an estimation of the CMS vector for the noisy speech. To estimate $\bar{\mathbf{m}}^{mfcc}$, the CMS vectors of the clean speech and noise, $\mathbf{m}^{mfcc}$ and $\tilde{\mathbf{m}}^{mfcc}$, are combined utilizing the PMC procedure (Veisi and Sameti, 2011). Due to the simplification assumptions in the domain transformation, the combination procedure given by PC-PMC method is not exact; it is an approximation called *log-normal* (Gales, 1995)

$$\begin{aligned} &\bar{\boldsymbol{\mu}}^{mfs}(i) = \log(\bar{\boldsymbol{\mu}}^{mel}(i)) - \frac{1}{2}\log\left(\frac{\bar{\boldsymbol{\Sigma}}^{mel}(i,i)}{(\bar{\boldsymbol{\mu}}^{mel}(i))^2} + 1\right); \\ &\bar{\boldsymbol{\Sigma}}^{mfs}(i,j) = \log\left(\frac{\bar{\boldsymbol{\Sigma}}^{mel}(i,j)}{\bar{\boldsymbol{\mu}}^{mel}(i) \cdot \bar{\boldsymbol{\mu}}^{mel}(j)} + 1\right);\ 1 \leqslant i,j \leqslant L_{mel}, \end{aligned} \tag{4}$$

$$\bar{\boldsymbol{\mu}}^{mfc} = \mathbf{C} \cdot \bar{\boldsymbol{\mu}}^{mfs} - \bar{\mathbf{m}}^{mfcc}; \quad \bar{\boldsymbol{\Sigma}}^{mfc} = \mathbf{C} \cdot \bar{\boldsymbol{\Sigma}}^{mfs} \cdot \mathbf{C}^{tr} \tag{5}$$

In the given procedure, each mixture of each state of the clean speech HMM is combined with all mixtures of all states of the noise HMM; therefore, the number of the states and mixtures for the noisy speech HMM $\bar{\lambda}^{mfc}$ are $\bar{M} = M \times \tilde{M}$ and $\bar{N} = N \times \tilde{N}$, respectively. The combination operator results in the following initial state probabilities, state transition probabilities and mixture weights for the noisy speech model:

$$\begin{aligned} &\bar{\pi}_{[\alpha_i\tilde{\alpha}_z]} = \pi_{[\alpha_i]} \cdot \tilde{\pi}_{[\tilde{\alpha}_z]}; \quad \bar{a}_{[\alpha_i\tilde{\alpha}_z][\alpha_j\tilde{\alpha}_x]} = a_{[\alpha_i\alpha_j]} \cdot \tilde{a}_{[\tilde{\alpha}_z\tilde{\alpha}_x]}; \\ &\qquad 1\leqslant i\leqslant M, 1\leqslant z\leqslant \tilde{M} \qquad\qquad 1\leqslant i,j\leqslant M, 1\leqslant z,x\leqslant \tilde{M} \\ &\bar{c}_{[\beta_k\tilde{\beta}_v]|[\alpha_i\tilde{\alpha}_z]} = c_{\beta_k|\alpha_i} \cdot \tilde{c}_{\tilde{\beta}_v|\tilde{\alpha}_z}, \\ &1\leqslant i\leqslant M, 1\leqslant z\leqslant \tilde{M}; 1\leqslant k\leqslant N, 1\leqslant v\leqslant \tilde{N} \end{aligned} \tag{6}$$

where $\bar{\pi}_{[\alpha_i\tilde{\alpha}_z]}$ represents the initial state probability of the noisy speech state $[\alpha_i\tilde{\alpha}_z]$. The noisy speech state $[\alpha_i\tilde{\alpha}_z]$ is resulted from the combination of the ith clean speech state, α_i, and the zth noise state, $\tilde{\alpha}_z$. $\bar{a}_{[\alpha_i\tilde{\alpha}_z][\alpha_j\tilde{\alpha}_x]}$ denotes the transition probability of transition from state $[\alpha_i\tilde{\alpha}_z]$ to state $[\alpha_j\tilde{\alpha}_x]$. Moreover, $\bar{c}_{[\beta_k\tilde{\beta}_v]|[\alpha_i\tilde{\alpha}_z]}$ indicates the mixture coefficient for the mixture component $[\beta_k\tilde{\beta}_v]$ of the state $[\alpha_i\tilde{\alpha}_z]$ where the noisy speech mixture $[\beta_k\beta_v]$ is constructed from the kth clean speech mixture β_k of the state α_i, and the vth noise mixture $\tilde{\beta}_v$ of the state $\tilde{\alpha}_z$.

To do the modeling in the MFS domain, the parameters of noise model $\tilde{\lambda}^{mfs}$ and the clean speech model λ^{mfs} in the MFS domain are combined to estimate the MFS domain parameters of noisy speech model, $\bar{\lambda}^{mfs}$. To this end, the same procedure is applicable but the operations of Eqs. (1) and (5) are no longer required.

2.2. Clean speech estimation

Let $f(\cdot)$ be a function such as $DFT(\cdot)$ or $|DFT(\cdot)|$ applied to $\mathbf{s}_t$. Given a noisy speech frame $\mathbf{y}_t \in \Re^L$, the speech enhancement system is designed to estimate $f(\hat{\mathbf{s}}_t)$ which minimizes an average distortion $E\{\Lambda(f(\hat{\mathbf{s}}_t), f(\mathbf{s}_t))\}$, where $\Lambda(\cdot)$ defines a distortion measure and $E\{\cdot\}$ denotes the mathematical expectation. Euclidean norm given as $\Lambda(f(\hat{\mathbf{s}}_t), f(\mathbf{s}_t)) \triangleq \|f(\hat{\mathbf{s}}_t) - f(\mathbf{s}_t)\|$ where $\|\mathbf{x}_t\| = \mathrm{sqrt}\sum_{i=0}^{L-1}\mathbf{x}_t(i)^2$, is a popular distortion measure that the minimization of the expected value of which results in the MMSE estimator (Ephraim, 1992b). This estimator is shown in Eq. (7) where the domain index (MFC or MFS) is dropped for brevity.

$$\begin{aligned} f(\hat{\mathbf{s}}_t) = E\{f(\mathbf{s}_t)|\mathbf{y}_{0:t}\} &= \sum_{i=1}^{\bar{M}}\sum_{k=1}^{\bar{N}} p_{\bar{\lambda}}(\bar{q}_t = \bar{\alpha}_i, \bar{u}_t \\ &= \bar{\beta}_k|\mathbf{y}_{0:t}) \cdot E\{f(\mathbf{s}_t)|\mathbf{y}_t, \bar{q}_t = \bar{\alpha}_i, \bar{u}_t = \bar{\beta}_k\} \end{aligned} \tag{7}$$

In this equation, $\mathbf{y}_{0:t} \triangleq \{\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_t\}$ denotes the noisy speech signal from time zero up to time t and $p_{\bar{\lambda}}$ defines the pdf of the noisy model. $\bar{q}_t$ and $\bar{u}_t$ denote the state and mixture of the noisy speech HMM, respectively, at time t, and $\bar{\alpha}_i$, $1 \leqslant i \leqslant \bar{M}$ and $\bar{\beta}_k$, $1 \leqslant k \leqslant \bar{N}$ indicate the state and mixture indexes, respectively. The MMSE estimator of Eq. (7) is the weighted sum of $\bar{M} \times \bar{N}$ individual MMSE estimators where the weights are the probabilities of choosing the individual estimators for the given noisy signal, (i.e., the filter weights, $p_{\bar{\lambda}}(\bar{q}_i, \bar{u}_t|\mathbf{y}_{0:t})$), and the individual MMSE estimators (i.e., the filter values), are defined as $E\{f(\mathbf{s}_t)|\mathbf{y}_t, \bar{q}_t, \bar{u}_t\}$). The conditional probability $p_{\bar{\lambda}}(\bar{q}_i, \bar{u}_t|\mathbf{y}_{0:t})$ is the posterior probability of the noisy speech model given $\mathbf{y}_{0:t}$ and can be efficiently calculated using 'forward–backward' algorithm (Ephraim, 1992c). However, to calculate $E\{f(\mathbf{s}_t)|\mathbf{y}_t, \bar{q}_t, \bar{u}_t\}$, the exact definition of $f(\cdot)$ is required. If we define $f(\mathbf{s}_t) = f_{sp}(\mathbf{s}_t) = DFT(\mathbf{s}_t) = \mathbf{s}_t^{spec}$, the evaluation of $E\{f(\mathbf{s}_t)|\mathbf{y}_t, \bar{q}_t, \bar{u}_t\}$ for the ith component of $f_{sp}(\hat{\mathbf{s}}_t)$ results in the following Wiener estimator (Lim and Oppenheim, 1978)

$$\begin{aligned} f_{sp}(\hat{\mathbf{s}}_t)(i) &= \frac{\mathbf{P}_{u_t|q_t}(i)}{\mathbf{P}_{u_t|q_t}(i) + \mathbf{P}_{\tilde{u}_t|\tilde{q}_t}(i)}\mathbf{y}_t^{spec}(i) \\ &= \left(\frac{\xi_{\bar{u}_t|\bar{q}_t}(i)}{\xi_{\bar{u}_t|\bar{q}_t}(i) + 1}\right) \cdot \mathbf{y}_t^{spec}(i) = \mathbf{H}_{\bar{u}_t|\bar{q}_t}^{sp}(i) \cdot \mathbf{y}_t^{spec}(i) \end{aligned} \tag{8}$$

In this estimator, $\mathbf{P}_{u_t|q_t} \triangleq E\{|\mathbf{s}_t^{spec}|^2\} \simeq (\boldsymbol{\mu}_{u_t|q_t}^{mag}(i))^2$, $1 \leqslant i \leqslant L_{mag}$ defines the power spectral density of the clean speech for the mixture component $[u_t = \beta_k | q_t = \alpha_i]$. Similarly $\mathbf{P}_{\tilde{u}_t|\tilde{q}_t}(i) \simeq (\tilde{\boldsymbol{\mu}}_{\tilde{u}_t|\tilde{q}_t}^{mag}(i))^2$, $1 \leqslant i \leqslant L_{mag}$ denotes the ith component of the power spectral density of noise mixture component $[\tilde{u}_t = \tilde{\beta}_k | \tilde{q}_t = \tilde{\alpha}_i]$ .The variables $\boldsymbol{\mu}_{u_t|q_t}^{mag}$ and $\tilde{\boldsymbol{\mu}}_{\tilde{u}_t|\tilde{q}_t}^{mag}$ indicate the mean parameters of the speech and noise HMMs, respectively, at the given state-mixture pairs in the MAG domain. To estimate $\boldsymbol{\mu}_{u_t|q_t}^{mag}$ and $\tilde{\boldsymbol{\mu}}_{\tilde{u}_t|\tilde{q}_t}^{mag}$ the corresponding MEL domain parameters of Eq. (2) are transformed as $\boldsymbol{\mu}^{mag} = \mathbf{MEL}^{-1} \cdot \boldsymbol{\mu}^{mel}$ in which, $\mathbf{MEL}^{-1}$ is the pseudo inverse of the Mel-scaled filter-bank transform $\mathbf{MEL}_{(L_{spec} \times L_{mel})}$, where L_{spec} denotes the number of spectrum coefficients and L_{mel} indicates the number of Mel-scaled filter banks. The auxiliary variable $\xi_{\bar{u}_t|\bar{q}_t}$ in Eq. (8) identifies *a priori signal-to-noise ratio* (*SNR*) (Ephraim and Malah, 1984) which its ith element is defined as $\xi_{\bar{u}_t|\bar{q}_t}(i) \triangleq \mathbf{P}_{u_t|q_t}(i)/\mathbf{P}_{\tilde{u}_t|\tilde{q}_t}(i)$. Furthermore, we define *filter function* as $\mathbf{H}(i) = \hat{\mathbf{s}}_t^{mag}(i)/\mathbf{y}_t^{mag}(i)$ where $\hat{\mathbf{s}}_t^{mag}(i)$ denotes the ith component of the estimated magnitude spectrum of $\mathbf{s}_t$. As shown in Eq. (8), the *filter function* of the Wiener estimation (i.e., the Wiener filter) at the ith frequency bin is defined as $\mathbf{H}^{sp}(i) = \xi(i)/(\xi(i)+1)$.

There are functions other than $f_{sp}(\mathbf{s}_t)$ used in the speech enhancement, too. In (Ephraim and Malah, 1984) an MMSE-based estimator was derived for $f_{spa}(\mathbf{s}_t) = |DFT(\mathbf{s}_t)| = \mathbf{s}_t^{mag}$ known as short-time spectral amplitude (STSA) estimator. The MMSE estimator for $f_{lsp}(\mathbf{s}_t) = \log|DFT(\mathbf{s}_t)| = \log(\mathbf{s}_t^{mag})$ that is called log-spectral amplitude (LSA) estimator, was given in (Ephraim and Malah, 1985), as well. The MMSE estimator for the power spectrum, i.e., $f_{psp}(\mathbf{s}_t) = |DFT(\mathbf{s}_t)|^2 = (\mathbf{s}_t^{mag})^2$, was given in (Wolfe and Godsill, 2003); and more generally, the MMSE estimator for the pth power of spectrum was obtained in (You et al., 2003). The performance of the noise reduction methods employing these estimators has been shown to be superior to the Wiener estimator in the non-model-based speech enhancement framework (Ephraim and Malah, 1984, 1985; Wolfe and Godsill, 2003). This fact motivates us to incorporate these techniques in the HMM-based speech enhancement framework as well. The *filter function*s corresponding to the mentioned functions are given as in Eqs. (9)–(11). In these equations, γ denotes the *posterior SNR* which its ith element is defined as $\gamma(i) = (\mathbf{y}_t^{mag}(i))^2/E\{(\mathbf{d}_t^{mag}(i))^2\}$, and $\chi(i)$ is defined as $\chi(i) = \gamma(i) \cdot \xi(i)/(\xi(i)+1)$. The function $I_i(\cdot)$ denotes the modified Bessel function of order i. Various approaches such as decision-directed method (Ephraim and Malah, 1984) were proposed to estimate ξ and γ parameters in the non-model-based speech enhancement framework. In the HMM-based speech enhancement system, these parameters are estimated based on the statistics of the speech and noise HMMs as in Eq. (12)

$$f_{spa}(\mathbf{s}_t) = |DFT(\mathbf{s}_t)| = \mathbf{s}_t^{mag} \Rightarrow \mathbf{H}^{spa}(i) = \frac{\sqrt{\pi\chi(i)}}{2\gamma(i)} \exp\left(-\frac{\chi(i)}{2}\right) \cdot \left[(1+\chi(i)) \cdot I_0\left(\frac{\chi(i)}{2}\right) + \chi(i) \cdot I_1\left(\frac{\chi(i)}{2}\right)\right] \quad (9)$$

$$f_{lsp}(\mathbf{s}_t) = \log|DFT(\mathbf{s}_t)| = \log(\mathbf{s}_t^{mag}) \Rightarrow \mathbf{H}^{lsp}(i) = \frac{\xi(i)}{\xi(i)+1} \exp\left\{\frac{1}{2}\int_{\chi(i)}^{\infty} \frac{e^{-t}}{t} dt\right\} \quad (10)$$

$$f_{psp}(\mathbf{s}_t) = |DFT(\mathbf{s}_t)|^2 = (\mathbf{s}_t^{mag})^2 \Rightarrow \mathbf{H}^{psp}(i) = \frac{\xi(i)}{\xi(i)+1}\left(\frac{\chi(i)+1}{\gamma(i)}\right) \quad (11)$$

$$\xi_{\bar{u}_t|\bar{q}_t}(i) = \left(\frac{\boldsymbol{\mu}_{u_t|q_t}^{mag}(i)}{\tilde{\boldsymbol{\mu}}_{\tilde{u}_t|\tilde{q}_t}^{mag}(i)}\right)^2; \quad \gamma_{\bar{u}_t|\bar{q}_t}(i) = \left(\frac{\mathbf{y}_t^{mag}(i)}{\tilde{\boldsymbol{\mu}}_{\tilde{u}_t|\tilde{q}_t}^{mag}(i)}\right)^2. \quad (12)$$

Practically, it is known that instead of applying Eq. (12) directly, it is beneficial to substitute the *instantaneous* magnitude spectrum $\mathbf{y}_t^{mag}$ with the *smoothed* form $\bar{\mathbf{y}}_t^{mag} = \rho_y \cdot \mathbf{y}_t^{mag} + (1-\rho_y) \cdot \bar{\boldsymbol{\mu}}_{\bar{u}_t|\bar{q}_t}^{mag}$ where ρ_y is the smoothing factor. Similarly, the $\boldsymbol{\mu}_{u_t|q_t}^{mag}$ can be substituted by $\hat{\mathbf{s}}_t^{mag}$ or $\bar{\mathbf{s}}_t^{mag} = \rho_s \cdot \hat{\mathbf{s}}_t^{mag} + (1-\rho_s) \cdot \boldsymbol{\mu}_{u_t|q_t}^{mag}$ where $\hat{\mathbf{s}}_t^{mag}$ denotes the estimation of $\mathbf{s}_t^{mag}$. Based on numbers of experiments for different values of ρ_y and ρ_s, we observed that the smoothed form did not provide significant improvement in our experiments. Therefore, we have used the instantaneous form in the evaluation of this paper in Section 4.

2.3. Approximation of estimators by hard decisioning

The filtering of noisy speech signal $\mathbf{y}_t$ in Eq. (7) is performed by incorporating all the mixture components of the noisy model $\bar{\lambda}$. This yields a time-consuming process especially when the numbers of mixtures are increased. To weaken the effect of the problem, an approximation of the filtering can be performed using only a single dominant state. Given a sequence of the noisy speech frames, the estimation of the corresponding most likely state sequence is achieved by applying the Viterbi algorithm (Rabiner, 1989). In addition, there is often more than one mixture component per state; therefore, the filtering of the noisy signal using the dominant state can be performed utilizing all mixtures in the state or using only the dominant mixture in the state (the most likely filter among all filters). These two cases result in the estimators given in Eqs. (13) and (14), respectively. In these equations, $\bar{\alpha}^*$ indicates the dominant state of the noisy model $\bar{\lambda}$ at time t given $\mathbf{y}_t$, and $\bar{\beta}^*$ denotes the dominant mixture component in the dominant state.

$$f(\hat{\mathbf{s}}_t) = \sum_{k=1}^{\bar{N}} p_{\bar{\lambda}}(\bar{q}_t = \bar{\alpha}^*, \bar{u}_t = \bar{\beta}_k | \mathbf{y}_{0:t}) \cdot E\{f(\mathbf{s}_t) | \mathbf{y}_t, \bar{q}_t = \bar{\alpha}^*, \bar{u}_t = \bar{\beta}_k\}, \quad (13)$$

$$f(\hat{\mathbf{s}}_t) = E\{f(\mathbf{s}_t) | \mathbf{y}_t, \bar{q}_t = \bar{\alpha}^*, \bar{u}_t = \bar{\beta}^*\}. \quad (14)$$

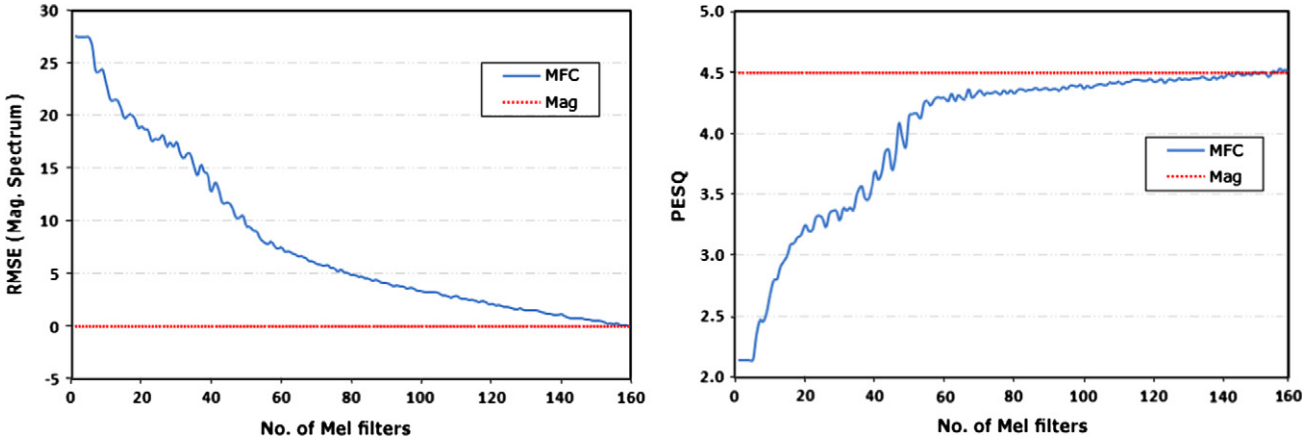
3. Parallel cepstral and spectral modeling

It has been proven that the performances of the ASR systems using the cepstral features is superior to systems employing other feature extraction methods (Gales, 1995; Veisi and Sameti, 2011). This superiority motivates us to employ cepstral features in the speech enhancement

framework. Generally, the selection of the filters in an HMM-based speech enhancement system, i.e. the calculation of the filter weights in Eq. (7), can be viewed as a recognition problem in which higher accurate recognition performance results in higher noise reduction. Nevertheless, the studies have indicated that the MFCC features are not necessarily the best parameters for recognition when dealing with noisy speech (Gales, 1995; Yu et al., 2008). To reduce the effect of this problem we have normalized the MFCC features using CMS method. It was shown (Veisi and Sameti, 2011) that the CMS normalization improves the speech recognition rate in noisy conditions. As shown in Fig. 1, the CMS normalized MFCC feature vector of a time-domain windowed frame $\mathbf{x}_t \in \Re^L$ can be generally defined as $\mathbf{x}_t^{mfc} = CMS(DCT\{LOG(MEL|DFT(\mathbf{x}_t)|)\})$. Considering only a symmetric half of the DFT coefficients, the DFT length is reduced to $L_{spec} = L/2 + 1$. Assuming the availability of the phase information, the inversion of *DFT* and magnitude operations are straight forward without loss of information. *MEL* is an $L_{spec} \times L_{mel}$ lossy transform due to the fact that $L_{spec} \gg L_{mel}$ in practice. The *LOG* operation and its inversion are lossless operators and we have $L_{mel} = L_{mfs.}$ *DCT* is an $L_{mfs} \times L_{mfc}$ matrix where $L_{mfs} \geqslant L_{mfc}$ (often $L_{mfc} = \lceil L_{mfs}/2 \rceil$ in practice) and it is often a lossy transform. The standard *CMS* is calculated dynamically and is not invertible. However, as described in the previous section, in this paper we have applied a fixed value for all signal frames as an approximation of the CMS vectors. This CMS operation is lossless. Therefore, the estimation of $\mathbf{x}_t$ from $\mathbf{x}_t^{mfc}$ which is required in the HMM-based speech enhancement described in Section 2, is an under-constrained problem and it is expected that the inversion from the MFC to the MAG domain introduces some distortion. To quantify the quality artifacts introduced by this inversion process, the MFC features are extracted for a time-domain clean speech signal from the TIMIT corpus, and then the time-domain signal is reconstructed from these features. The phase of the original signal is used in the reconstruction. In this experiment, the root mean square error (RMSE) between the original and the reconstructed magnitude spectrums, defined as $\mathrm{RMSE}^{mag} = \mathrm{sqrt}\sum_{t=0}^{T-1}\sum_{l=0}^{L_{spec}-1}(\mathbf{s}_t^{mag}(l) - \hat{s}_t^{mag}(l))^2$ for the signal with T frames, and the perceptual evaluation of speech quality (PESQ) (Perceptual Evaluation of Speech Quality (PESQ), 2001) measures are calculated and illustrated in Fig. 2 for different numbers of Mel-scaled filters. In this figure, the inversions from the MFC and magnitude spectrum domains are denoted by MFC and Mag, respectively. This experiment clearly confirms the degradation introduced by the inversion from the MFC features. In this experiment, it is assumed that $L_{mfs} = L_{mfc;}$ therefore the DCT matrix is invertible and the reconstruction from the MFC and MFS domains are the same. For $L_{mfc} < L_{mfs}$, the distortion from the MFC is higher than the MFS. Similar results were achieved for the RMSE of the waveform and the SNR measures.

Generally, as DFT coefficients are modified, the direct inversion of the DFT coefficients using standard overlap-and-add reconstruction is no longer valid. There are several methods in (Griffin and Lim, 1984; Imai, 1983) that are proposed to obtain a valid spectrum and generate a reconstructed waveform that is close enough to the original waveform. In (Griffin and Lim, 1984), two methods called inverse short-time Fourier transform (STFT) and inverse STFT magnitude, are presented that are optimum in terms of least squared error (LSE). In (Imai, 1983), Mel log spectrum approximation (MLSA) filter is derived to synthesize the speech signal from Mel log-spectrum which is commonly used in model-based speech synthesis. Furthermore, there are other techniques in (Tokuda et al., 2000, 1995) that are proposed for speech parameter generation in HMM-based speech synthesis applications.

In this paper, a parallel cepstral and spectral modeling (PCS) is proposed in order to either utilize the benefits of the MFC/MFS features in the HMM-based speech enhancement system and to eliminate the distortion effect of the inversion process in calculating the values of filters.



(a) Root mean square error of magnitude spectrum (b) PESQ

Fig. 2. Distortion introduced by the inversion from the MFC domain versus the number of Mel filters.

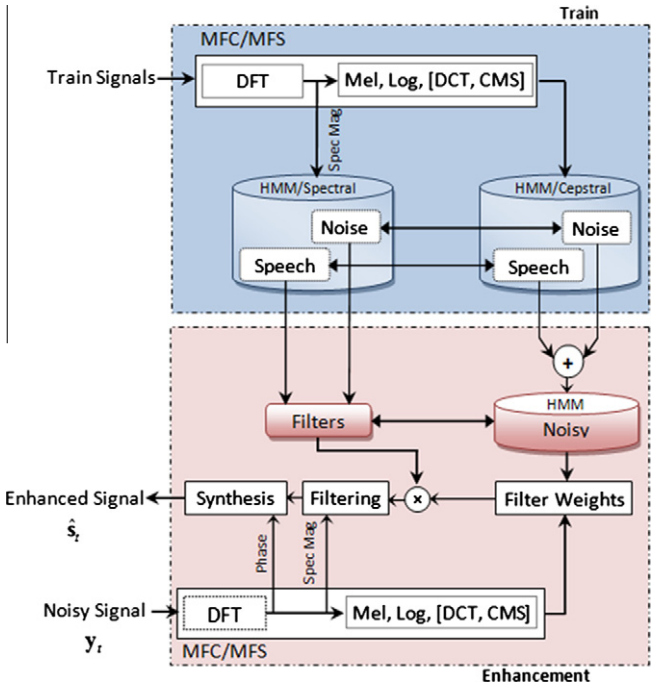


Fig. 3. Parallel cepstral and spectral modeling for HMM-based MMSE speech enhancement.

As shown in Fig. 3, in PCS method, the HMM modeling is performed concurrently in both magnitude spectral and cepstral domains. The modeling in the cepstral domain is performed using the conventional HMM formulation given in Section 2. In order to create the concurrent magnitude spectral models, it is assumed that the alignment of the frames and HMM states in the cepstral and magnitude spectral domains are identical. If we define $\zeta_t^{mfc}(\alpha,\beta)$ as the probability of being in mixture β of stateα given the observation sequence $\mathbf{s}_{0:t}^{mfc}=\{\mathbf{s}_0^{mfc},\ldots,\mathbf{s}_t^{mfc}\}$ in the MFC domain, then the re-estimation of the mean vector and covariance matrix parameters are performed as Eqs. (15) and (16), respectively; where $\mathbf{s}_t^{mag}$ is the observation vector in the magnitude spectral domain corresponding to $\mathbf{s}_t^{mfc}$, and T indicates the total number of observations. Other HMM parameters of λ^{mag} are identical to the equivalent parameters of λ^{mfc}. A similar model in the MAG domain is also estimated for the noise, $\tilde{\lambda}^{mag}$.

In the enhancement phase, the individual MMSE estimators (i.e., the filter values) are calculated using the statistics of the magnitude spectral domain speech and noise models, λ^{mag} and $\tilde{\lambda}^{mag}$, i.e., $\ddot{\mathbf{H}}_{\bar{u}_t|\bar{q}_t}^{sp}(i)=(\ddot{\boldsymbol{\mu}}_{u_t|q_t}^{mag}(i))^2/[(\ddot{\boldsymbol{\mu}}_{u_t|q_t}^{mag}(i))^2+(\ddot{\tilde{\mu}}_{\bar{u}_t|\bar{q}_t}^{mag}(i))^2]$ for the Wiener filter. On the other hand, the weights of the filters, i.e., the conditional probabilities $p_{\bar{\lambda}^{mfc}}(\bar{q}_t,\bar{u}_t|\mathbf{y}_{0:t})$, are calculated using the noisy speech HMM in the cepstral domain, $\bar{\lambda}^{mfc}$. The PCS modeling is applicable to both MFC and MFS domains.

$$\ddot{\boldsymbol{\mu}}_{\beta|\alpha}^{mag}=\frac{\sum_{t=0}^{T-1}\zeta_t^{mfc}(\alpha,\beta)\cdot\mathbf{s}_t^{mag}}{\sum_{t=0}^{T-1}\zeta_t^{mfc}(\alpha,\beta)} \tag{15}$$

$$\ddot{\boldsymbol{\Sigma}}_{\beta|\alpha}^{mag}=\frac{\sum_{t=0}^{T-1}\zeta_t^{mfc}(\alpha,\beta)\cdot\left(\mathbf{s}_t^{mag}-\ddot{\boldsymbol{\mu}}_{\beta|\alpha}^{mag}\right)\cdot\left(\mathbf{s}_t^{mag}-\ddot{\boldsymbol{\mu}}_{\beta|\alpha}^{mag}\right)^{tr}}{\sum_{t=0}^{T-1}\zeta_t^{mfc}(\alpha,\beta)} \tag{16}$$

To demonstrate the effectiveness of the PCS modeling, the *filter distance* of Eq. (17) is defined that calculates root mean square log spectral distance (RMS-LSD) between the *ideal filter*, $\mathbf{H}^{idl}$, and another filter for the filtering of the noisy speech $\mathbf{y}_{0:T-1}=[\mathbf{y}_0,\ldots,\mathbf{y}_t,\ldots,\mathbf{y}_{T-1}]$. The lth component of the ideal Wiener filter at time t corresponding to $\mathbf{y}_t$ is defined as $\mathbf{H}_t^{idl}(l)=(\mathbf{s}_t^{mag}(l))^2/[(\mathbf{s}_t^{mag}(l))^2+(\mathbf{d}_t^{mag}(l))^2]$. To calculate the elements of $\mathbf{H}^{idl}$ in the filtering of a noisy speech frame $\mathbf{y}_t=\mathbf{s}_t+\mathbf{d}_t$, the original magnitude spectrum of the corresponding clean speech signal $\mathbf{s}_t$ and noise signal $\mathbf{d}_t$ are used.

$$\Delta=\mathrm{sqrt}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{l=0}^{L-1}\left(\log\frac{\mathbf{H}_t(l)}{\mathbf{H}_t^{idl}(l)}\right)^2. \tag{17}$$

The distance between the ideal Wiener filter and non-ideal Wiener filters are calculated and the values of the distances for different noise types and various SNR levels are shown in Fig. 4. The non-ideal Wiener filters used in this experiment include the filters estimated using the MFC and MFS domain HMMs (shown as *MFC* and *MFS* in the figure), and the PCS modeling using the MFC and MFS features (shown as *PCS–MFC* and *PCS–MFS*, respectively). These filters are calculated based on Eq. (7). For each noisy condition, the distance values are computed for (and averaged over) all noisy speech files of the test set

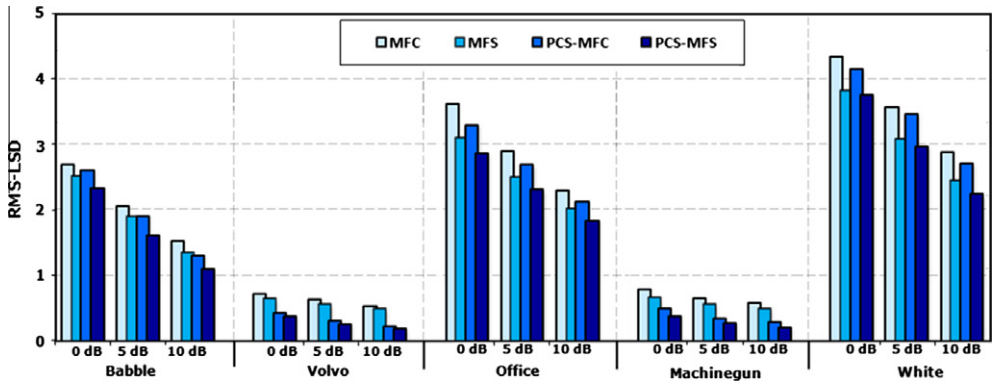


Fig. 4. Distance of ideal filter from the filters resulted using MFC, MFS, PCS–MFC and PCS–MFS modeling methods for different noisy conditions.

from TIMIT corpus introduced in Section 5. It is observed from Fig. 4 that the PCS modeling results in closer filters to the ideal filter than the non-PCS ones.

4. Experimental evaluations

In this section, we describe the experimental setup and present the results of our evaluations. In the experiments, speech data with the sampling rate of 16 kHz is selected from TIMIT corpus in two sets, train set and test set. The train set contains 300 sentences of clean speech (about 16 min) from 30 speakers (10 females and 20 males) where each speaker has read 10 sentences. The clean speech test set includes 10 sentences from 10 speakers (4 females and 6 males). The train and test sets are exclusive, i.e., there are no common sentences or common speakers between the two sets. Five noise types including white, office, babble, Volvo and machinegun, taken from the Noisex92 corpus (except the office noise), are added (biased to 1 dB lower than the target SNR level) to the clean speech test set in three SNR levels of 0, 5 and 10 dB. The power spectrum densities (PSD) of the noise types used in the evaluations are given in Fig. 5. It shows that white noise is a full-band noise while Volvo and machinegun are low frequency noises.

For feature extraction, the analysis is performed in blocks of 20 ms ($L = 160$) windowed using Hamming window with 50% overlap between the adjacent frames. The number of Mel-filters and cepstral coefficients are $L_{mel} = 23$ and $L_{mfc} = 12$, respectively. In all experiments, fully connected HMMs are used for both speech and noise. Each speech HMM consists of 16 states and 16 mixture components per state and each noise HMM consists of four states and four mixtures per state. The covariance matrix is assumed diagonal and full in the experiments using the MFC and MFS features, respectively. In feature extraction, only static features are used and dynamic features, i.e., delta and delta–delta, are not considered. The time-domain enhanced speech signal is reconstructed from the estimated magnitude and the phase information of the noisy (original) signal applying overlap-and-add method. In the experiments reported in this section, it is assumed that we have prior knowledge of the type of the noise environment such that the correct noise model is used in the enhancement process. This assumption has been used in (Ephraim et al., 1989; Ephraim, 1992a,–c; Sameti, 1994; Sameti et al., 1998; Logan, 1998; Zhao and Kleijn, 2007) as well and it is acceptable in some application. In many of real applications of the HMM-based speech enhancement systems, it is necessary to have a "noise identification" module that identifies the noise type before performing the enhancement.

The results of the enhancement systems are evaluated using objective and subjective tests. In the objective evaluation overall SNR (in dB) (Loizou, 2007), frequency weighted segmental SNR (in dB) (Loizou, 2007) and PESQ (Evaluation of Speech Quality (PESQ), 2001) criteria are used, and mean opinion score (MOS) test (Loizou, 2007) is used in the subjective evaluations. The overall SNR is indicated as *SNR* and the frequency weighted segmental SNR is denoted as SNR_{fws}. The overall SNR and the SNR of each frequency band are confined to the perceptually meaningful range between 35 and −10 dB. In calculation of SNR_{fws}, the frequency bands and their weights are computed as in (Loizou, 2007). The speech enhancement systems in the Mel-frequency domain that was described in Section 2 are referred to as *MFC–HMM* and *MFS–HMM* for Mel-frequency *cepstral* (MFC) coefficients and Mel-frequency *spectral* (MFS) coefficients, respectively. The PCS method introduced in Section 3 is denoted by *PCS–MFC* and *PCS–MFS* for the MFC and MFS features, respectively. The performances of the proposed methods are compared with the traditional renowned speech enhancement systems including MMSE-STSA (Ephraim and Malah, 1984), Log-MMSE (Ephraim and Malah, 1985), Wiener filtering based on wavelet thresholding (Hu et al., 2004), spectral subtraction (Berouti et al., 1979) and the improved AR-HMM based speech enhancement (Sameti et al., 1998). These methods are referred to as *MMSE*, *LogMMSE*, *SpecSub*, *Wiener* and *AR-HMM*,

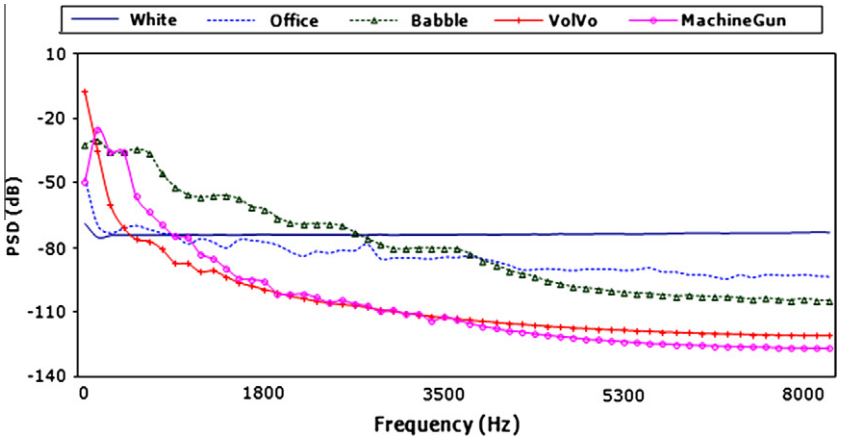


Fig. 5. Power spectrum densities (in dB) of the noise types used in the evaluations.

respectively. The values of the parameters in these methods are identical to the values given in the original references.

To estimate PSD of noise in the AR-HMM, we have used noise HMM as it was done in the original reference (Sameti et al., 1998). The training data for all noise types are same as the data we have used in the proposed methods. However, to estimate the noise PSD in the non-model-based methods (i.e., MMSE, LogMMSE, SpecSub and Wiener) it is assumed that first five frames of noisy signal are noise samples (it is a correct assumption in our test sets) and the spectrum of these frames are averaged and used as the initial estimate of noise PSD. For other frames of noisy signal, a voice activity detector (VAD) is used to detect the speech/non-speech frames and the noise PSD is updated for the non-speech (i.e., noise) frames using the well-known smoothed form: $\hat{\mathbf{d}}_t^{spec} = g \cdot \hat{\mathbf{d}}_{t-1}^{spec} + (1-g) \cdot \mathbf{y}_t^{spec}$, where the smoothing factor $g = 0.98$ is used in our experiments.

4.1. Performance comparison of various filter functions

The performances of the various filtering functions that were studied in Section 2.2 are evaluated using the HMM-based speech enhancement system with the MFC features. The comparative results of the Wiener filter, $\mathbf{H}^{sp}$, and non-Wiener filters, $\mathbf{H}^{spa}$ (defined in Eq. (9)) and $\mathbf{H}^{lsp}$ (defined in Eq. (10)) on the selected noise types for 0 and 10 dB input signals are given in Table 1. The last columns of the tables indicate the average values over three noise types and the given SNR levels. A slight improvement in *SNR* and *PESQ* is resulted using $\mathbf{H}^{spa}$ and $\mathbf{H}^{lsp}$ in comparison with the Wiener filter on the stationary white noise. Moreover, the results demonstrate a slight improvement by $\mathbf{H}^{lsp}$ in *SNR* and *PESQ* on the non-stationary babble noise and show the reduction of the performance on the Volvo noise, particularly in low SNR levels. The incorporation of the non-Wiener estimators in the HMM-based speech enhancement framework is motivated by the fact they are more effective than the Wiener filter in the non-model-based methods (Ephraim and Malah, 1984, 1985). Our results indicate that for the MFC domain HMM-based speech enhancement, non-Wiener filters do not perform better than the Wiener filter. Furthermore, non-Wiener filters are more complex and require higher computational cost. We did the experiments on 5 dB input signals and other noise types as well. Their results are not reported for brevity, but similar results were obtained.

Table 1
Comparative results of the MFC–HMM using different filter functions.

Filter type	Data						Average
	White		Babble		Volvo		
	0 dB	10 dB	0 dB	10 dB	0 dB	10 dB	
(a) Overall SNR (dB)							
No enhancement (noisy signal)	−1.0	9.0	−0.7	9.3	−1.0	9.0	4.08
$\mathbf{H}_{sp}$ (spectrum, i.e., Wiener)	7.7	14.6	5.3	12.4	16.1	20.3	12.74
$\mathbf{H}_{spa}$ (spectral amplitude)	7.8	14.7	5.0	12.4	12.6	20.6	12.16
$\mathbf{H}_{lsp}$ (log-spectral amplitude)	7.8	14.8	5.1	12.5	11.6	20.8	12.10
(b) PESQ							
No enhancement (noisy signal)	1.4	2.2	1.7	2.5	3.1	3.8	2.44
$\mathbf{H}_{sp}$ (spectrum, i.e., Wiener)	2.0	2.8	2.0	2.7	3.6	4.1	2.84
$\mathbf{H}_{spa}$ (spectral amplitude)	2.1	2.8	2.0	2.7	3.4	3.9	2.82
$\mathbf{H}_{lsp}$ (log-spectral amplitude)	2.2	2.8	2.1	2.7	3.4	3.9	2.85

Table 2
Comparative results of the MFC-HMM speech enhancement system using hard-decision filter selection

Num filter(s)	Data						Average
	White		Babble		Volvo		
	0 dB	10 dB	0 dB	10 dB	0 dB	10 dB	
(a) Overall SNR (dB)							
No enhancement (noisy signal)	−1.0	9.0	−0.7	9.3	−1.0	9.0	4.08
AllStates-AllMixs	7.7	14.6	5.3	12.4	16.1	20.3	12.74
OneState-AllMixs	7.5	14.3	5.4	12.2	16.5	20.5	12.72
OneState-OneMix	7.0	13.7	4.8	11.7	16.2	20.4	12.30
(b) PESQ							
No enhancement (noisy signal)	1.4	2.2	1.7	2.5	3.1	3.8	2.44
AllStates-AllMixs	1.98	2.75	1.96	2.72	3.56	4.05	2.84
OneState-AllMixs	1.93	2.71	1.81	2.70	3.58	4.04	2.80
OneState-OneMix	1.69	2.62	1.38	2.62	3.56	4.07	2.66

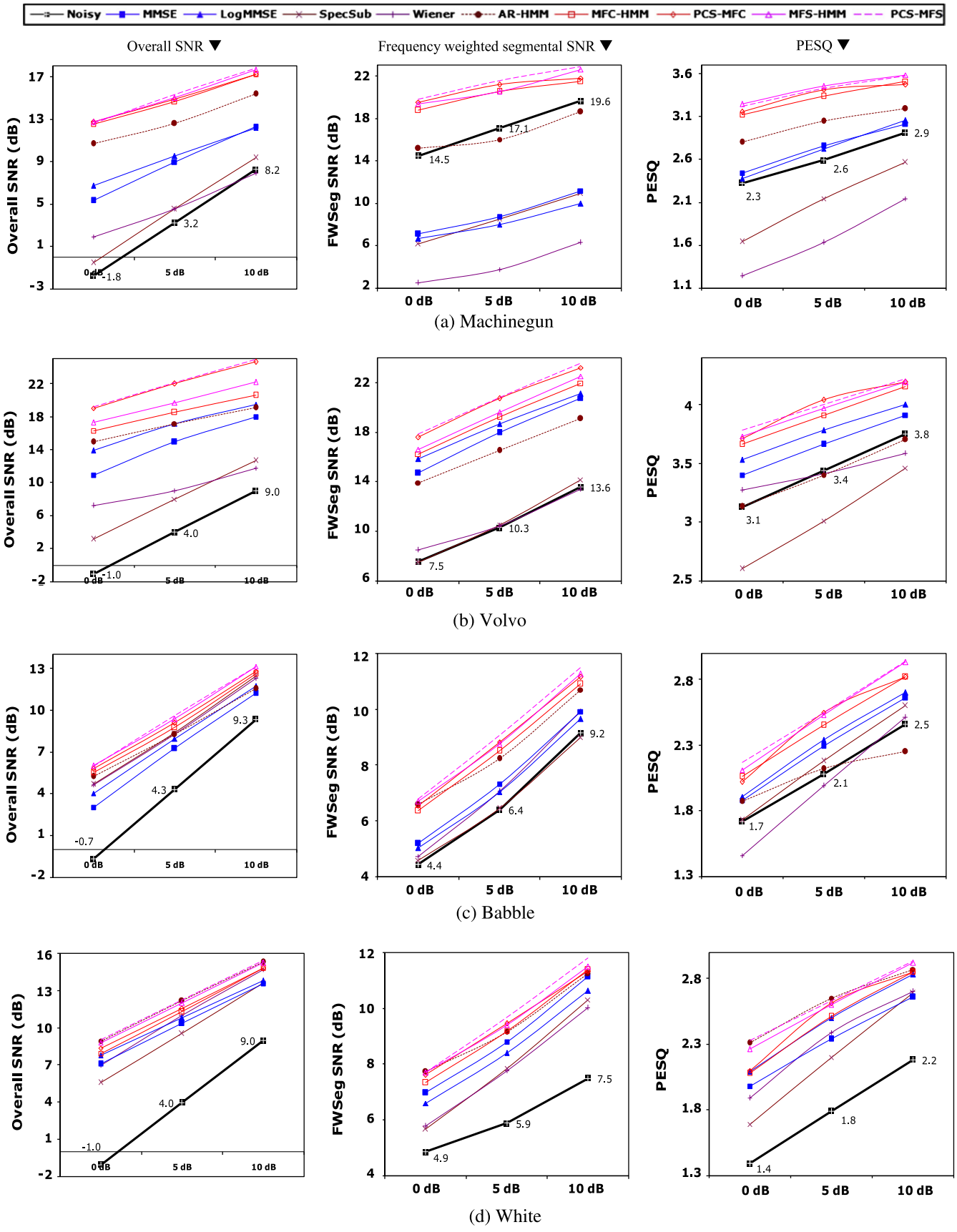
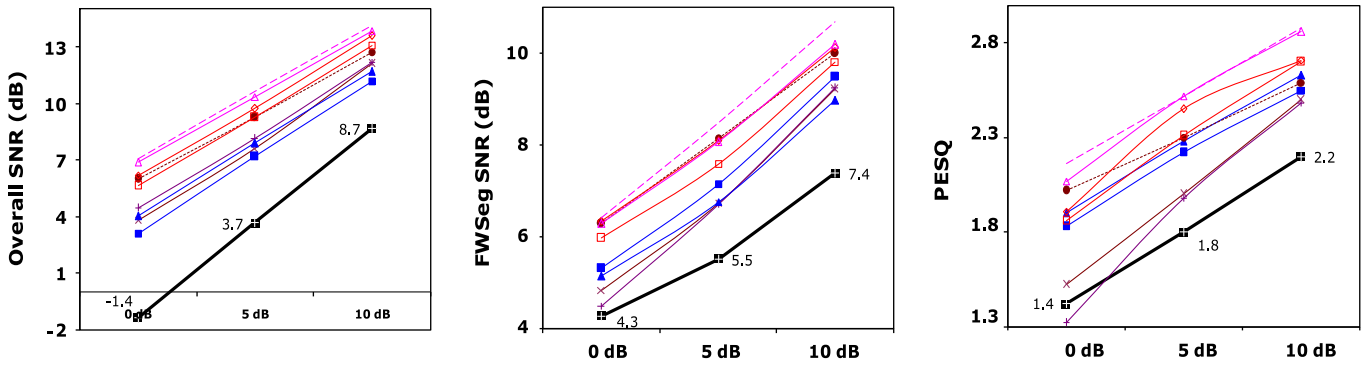


Fig. 6. Overall SNR, frequency weighted segmental SNR and PESQ for various noise types and SNR levels achieved by the proposed and reference methods.



(e) Office

Fig. 6. (*continued*)

4.2. Performance of the hard-decision filter selection

The evaluation of the MMSE estimator with hard-decision filter selection is examined using the MFC–HMM system. This approximation is beneficial since it results in lower computational cost in the speech estimation process. The evaluations are performed for three cases, (1) *AllStates-AllMixs*: all mixture components in all states are used to filter the noisy signal as shown in Eq. (7) (soft-decisioning), (2) *OneState-AllMixs:* all mixtures of the dominant state is used as shown in Eq. (13) and (3) *OneState-OneMix:* only the dominant mixture of the dominant state performs the filtering, as given in Eq. (14) (hard-decisioning). The results of these experiments for three noise types with two SNR levels are shown in Table 2 in terms of *SNR* (Table 2a) and *PESQ* (Table 2b). The last columns of the tables present the average values over three noise types and different SNR levels. According to the results, we can observe that the *AllStates-AllMixs* soft filtering results in higher performance than the two other cases. The *OneState-AllMixs* method results in a similar performance to the *AllStates-AllMixs*. The results given by the hard-decision filtering, *OneState-OneMix*, is lower than the other cases for white and babble noise types. Similar results are achieved for office noise and are not given here for the brevity. The performance of *OneState-OneMix* is interestingly similar to the other methods in the presence of the non-stationary Volvo and machinegun noises (the results of the machinegun noise are not shown for the brevity). These observations indicate that in the presence of some noise types, noisy speech is well clustered in separate classes.

The computational complexity of *AllStates-AllMixs*, *OneState-AllMixs* and *OneState-OneMix* cases for a noisy speech signal with T frames are in order of $O(T \times \bar{M} \times \bar{N})$, $O(T \times \bar{N}) + O(T \times \bar{M}^2)$ and $O(T) + O(T \times \bar{M}^2)$, respectively, where $O(T \times \bar{M}^2)$ indicates the complexity of the Viterbi algorithm. The complexity of *OneState-OneMix* is always lower than the complexity of the *OneState-AllMixs*; however, the orders of the computational costs of these methods depend on the number of the states and mixtures of the speech and noise HMMs.

4.3. Comparative evaluations

4.3.1. Objective evaluations

The results of evaluations using objective criteria are given in this section. The values of *SNR*, SNR_{fwS} and *PESQ* resulted by the proposed and the reference methods for five types of noises in three SNR levels are shown in Fig. 6. The results in this figure are calculated by averaging over 10 noisy test files. The values of the criteria for the noisy signal (i.e., the reference values) are also given in this figure. In the experiments of the proposed methods, we have applied Eq. (7) (all mixture components in all states) and have used Wiener filter, $\mathbf{H}^{sp}$, to estimate the clean speech. The results indicate that the Mel-frequency domain speech enhancement systems including MFC–HMM, MFS–HMM, PCS–MFC and PCS–MFS consistently outperform the reference methods for non-stationary noise types such as babble, Volvo and machinegun. Generally, PCS modeling improves the performance of the system for both MFC and MFS features. Due to the high correlation between PESQ scores as an objective measure and MOS subjective measure (Perceptual Evaluation of Speech Quality (PESQ), 2001), higher PESQ scores for the proposed methods denote better signal quality. This fact is verified by formal listening MOS tests that are given in the continuant of the paper.

From the results shown in Fig. 6, we could observe that the results of the five Mel-frequency domain methods in dealing with machinegun noise are consistently superior to the other methods for all conditions. Although the improvements achieved by these methods are close to each other, the PCS–MFS demonstrates slightly higher performances than the others. As given in Fig. 6(b), the PCS methods considerably outperform the other methods in the presence of the Volvo noise. The *PESQ* values are lower than the reference values (i.e., the *PESQ* of noisy signal) for a few cases such as in the presence of machinegun noise for Spec-Sub and Wiener methods, and in the presence of the Volvo noise for the Spec-Sub. The values of *PESQ* for the Wiener and AR-HMM techniques further decrease and become lower than the *PESQ* of noisy signal as the input SNR level of the Volvo noise increases. For the

non-stationary babble noise, the MFS domain methods (i.e., PCS–MFS and MFS–HMM) surpass the other methods in all conditions. In addition, it is observed that in the presence of this noise, the values of criteria resulted by AR-HMM rapidly decreases as the input signal SNR level is increased (the *PESQ* values becomes lower than the reference value for noisy signal of 10 dB). In the presence of stationary white noise (shown in Fig. 6(d)), the *SNR* values of AR-HMM and PCS–MFS are similar; however the Mel-frequency domain methods (especially PCS–MFS) result in better SNR_{fws} particularly for the higher input SNR levels. AR-HMM and PCS–MFS result in higher *PESQ* values especially in low SNR levels for the white noise. The results of the methods in the presence of office noise given in Fig. 6(e) show that the PCS–MFS outperform the others in all conditions and the other MFS-domain methods perform better than the remaining methods in terms of *SNR* and *PESQ*.

The efficiency of the proposed methods is also demonstrated from the spectrograms shown in Figs. 7 and 8 for

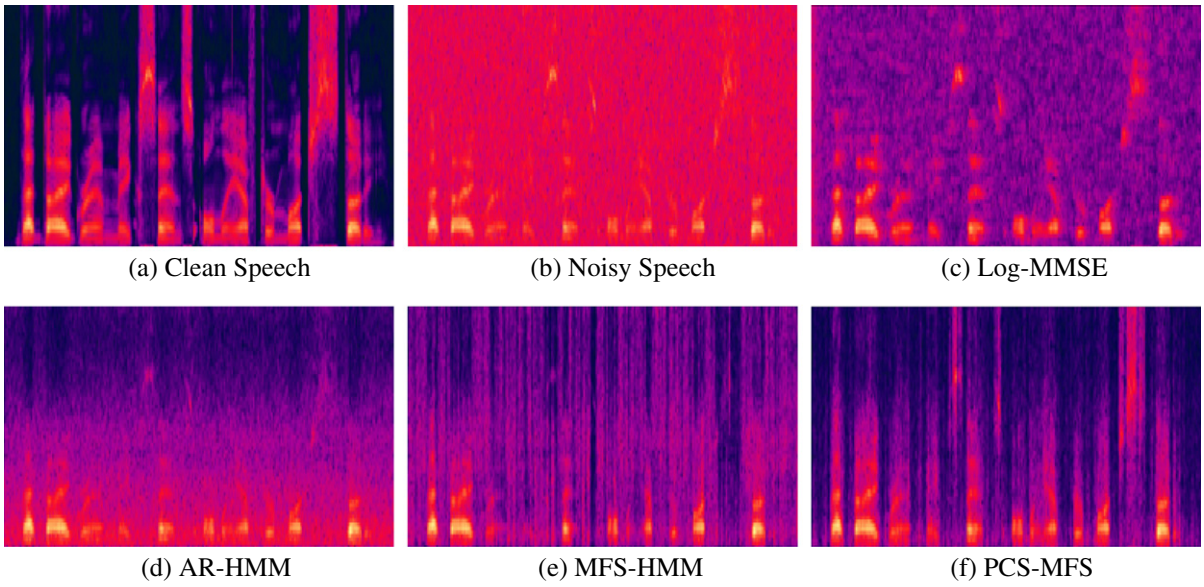


Fig. 7. Spectrogram of (a) a clean speech signal, (b) noisy speech signal at 0 dB by white noise and (c–f) enhanced speech signals for the selected proposed and reference methods.

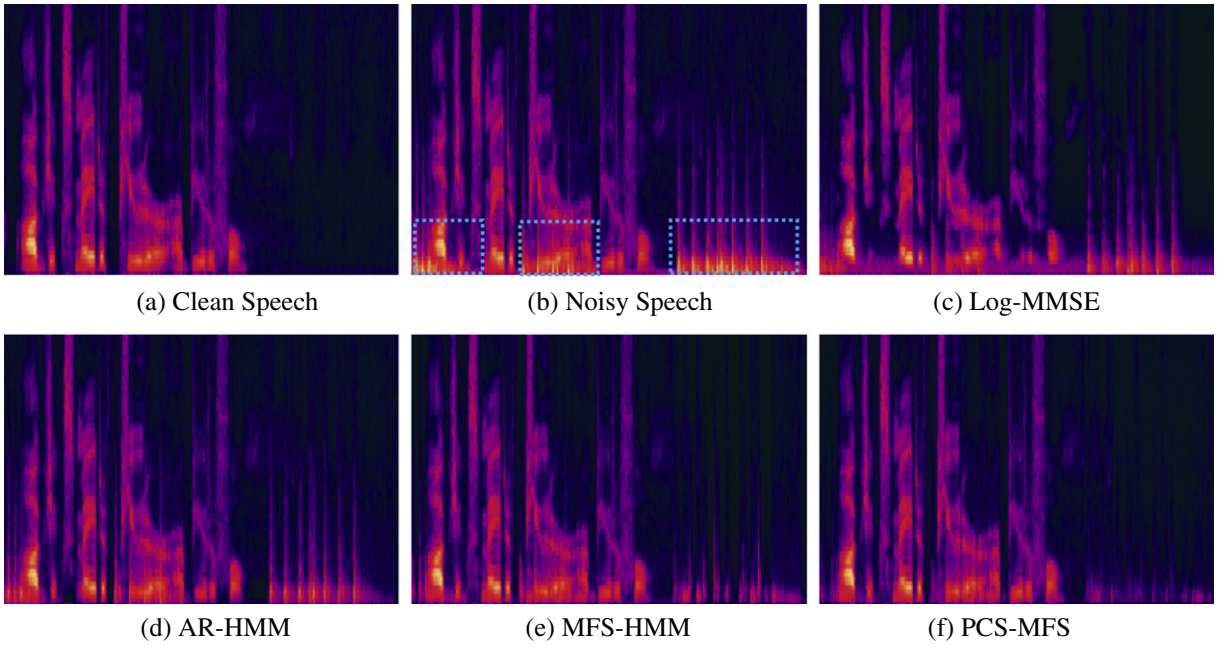
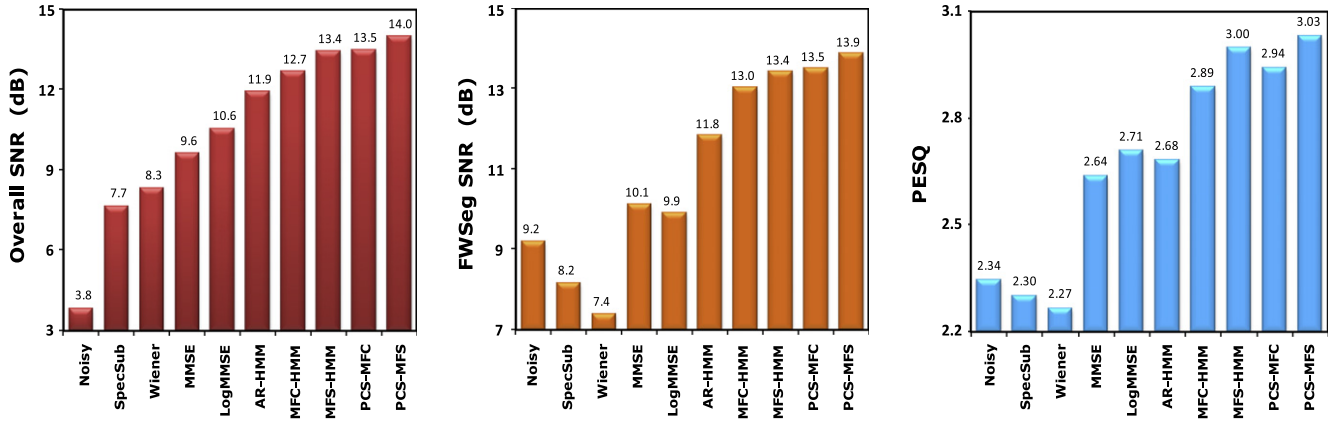


Fig. 8. Spectrogram of (a) a clean speech signal, (b) noisy speech signal at 0 dB by machinegun noise and (c–f) enhanced speech signals for the selected proposed and reference methods.



(a) Overall SNR (b) Frequency Weighted Segmental SNR (c) PESQ

Fig. 9. Average of (a) overall SNR, (b) frequency weighted segmental SNR and (c) PESQ, over all sentences in the noisy test sets including five noise types in three SNR levels.

white and machinegun noises, respectively, at 0 dB. These figures display the spectrograms of a clean, noisy and enhanced speech signal for the selected reference and proposed methods. As it is shown in Fig. 7, the noisy speech signal is totally masked by the noise. Although the Log-MMSE has enhanced the signal but the noise is remained in all frequencies and the background musical noise is generated. The musical noise of this method is evidence from the granularity of the spectrogram. The AR-HMM has resulted in better enhancement than Log-MMSE but the background noise is remained especially in lower frequencies. The proposed PCS–MFS method has produced the spectrogram that is closer to the clean spectrogram.

The higher performance of the proposed methods in dealing with the non-stationary machinegun noise is illustrated in Fig. 8. The burst power of the noise is shown using the dotted boxes in Fig. 8(b). As it can be viewed, the noise is not removed by the reference methods while the proposed methods, especially PCS–MFS, have almost eliminated the noise.

In addition to the detailed results given in Fig. 6, the average of the evaluation criteria over all noise types and all SNR levels are demonstrated in Fig. 9. From this figure that gives a general comparison between the proposed and the reference methods, the following observations can be conducted:

- The HMM-based techniques outperform the other approaches including non-model-based statistical methods such as MMSE and Log-MMSE. This is because the HMM-based methods employ a priori knowledge of the noise and speech signal and assumes that the true noise type is selected. However, this assumption limits the application of this approach unless a noise identification method is available.
- Higher improvements are achieved in the Mel-frequency domain (i.e., MFC and MFS). This is due to the fact that filter selection is performed more accurately in the Mel-frequency domain than in the spectrum domain.
- The proposed PCS modeling improves the performance of the enhancement system, i.e., the PCS–MFC is better than the MFC–HMM and the PCS–MFS is superior to the MFS–HMM for all testing conditions. From the results, the average values of *SNR*, SNR_{fws} and *PESQ* for the PCS–MFS method are the highest. This happens because the PCS modeling provides accurate values for the suppression filters.
- The MFS features yield better performance than the MFC coefficients in the HMM-based speech enhancement system. This is because the discrete cosine transform is not performed in the extraction of the MFS features.
- The results of SpecSub and Wiener methods show that they have improved the *SNR* but have decreased the SNR_{fws} and *PESQ* values.

4.3.2. Evaluation of non-model-based methods in the presence of known noise type

As it was mentioned, in the evaluations of the model-based methods it is assumed that the noise type is known. To have a fair comparison between the model-based methods (i.e., the proposed methods and AR-HMM) and other non-model-based reference methods we have evaluated the later methods assuming that the noise type is given. In this evaluation, the noise PSD is estimated from a known noise signal and it is used as an initial estimate of the PSD. However, one may update the initial PSD during the enhancement based on a VAD decision. We have evaluated the non-model-based methods in both cases i.e., (1) only the initial estimation of the noise PSD is used during the enhancement process and the noise PSD is not updated, and (2) the initial estimation of noise PSD is updated during the enhancement based the VAD decision. Our evaluations show that the performance of the second case is higher than the first case. For the second case, average values of the evaluation criteria over all noisy conditions (i.e., all noise types at three SNR levels) are given in Table 3. As it is observed,

Table 3
Results of the non-model-based speech enhancement methods assuming the noise type is known.

Method	Criteria		
	SNR	SNR_{fws}	*PESQ*
Noisy	3.8	9.2	2.34
MMSE	9.6	10.1	2.64
MMSE (known noise)	8.4	10.2	2.64
LogMMSE	10.6	9.9	2.71
LogMMSE (known noise)	9.3	10.2	2.68
SpecSub	7.7	8.2	2.30
SpecSub (known noise)	8.4	10.1	2.55
Wiener	8.3	7.4	2.27
Wiener (known noise)	8.1	7.6	2.23

using the known noise type has improved the performance of the SpecSub method. The improvement is achieved in most noisy conditions and almost for all criteria. However, the improvement is still lower than the model-based methods. Also, as it is shown, the average values of SNR_{fws} are marginally improved that come from the improvement of this criterion for the machinegun noise. Furthermore, the known noise type experiment results in higher *PESQ* values for white noise. For all other cases, the performance is decreased using the known noise type assumption.

The lower performances of the known noise type evaluations come from the fact that the estimated PSD from the known noise signal is only an average of the noise PSD and does not include the rapid variations of the noise. Therefore, when this average is also smoothed using the update formula, i.e., $\hat{\mathbf{d}}_t^{spec} = g \cdot \hat{\mathbf{d}}_{t-1}^{spec} + (1-g) \cdot \mathbf{y}_t^{spec}$, it cannot handle the dynamics of the noise quickly. On the other hand, in the model-based methods and especially in HMM, the dynamic variations of the noise are also modeled. Generally, the known noise type assumption only affects the performance of the model-based methods and cannot improve the noise reduction ability of the non-model-based techniques because the non-model-based methods do not have a mechanism of using this 'prior' information. The higher performance of the SpecSub method in the given experiment is probably due to the fact we have used a simple VAD in its implementation and the new estimation of PSD is more accurate. In addition, the higher *PESQ* values for the stationary white noise are because that the new estimate of PSD for this noise results in an accurate average of PSD and as the noise is stationary, it has not rapid changes during the enhancement.

4.3.3. Subjective evaluations

The objective evaluation in the previous sections shows the superiority of HMM-based speech enhancement in MFC and MFS domains. In this section, we demonstrate the results of subjective quality assessment using mean opinion score (MOS) test for selected proposed and reference techniques. The reference methods that are selected in the evaluation of this experiment have resulted in higher performance than the other methods in the objective evaluations. The evaluations are done on five noise types at input SNR levels of 0 and 10 dB. The speech enhancement systems are scored by 10 speakers using the scoring criterion established in MOS test from 1 (bad quality, very annoying) to 5 (excellent quality) (Loizou, 2007). The average values of MOS scores for the selected methods are given in Fig. 10. The average values of the MOS scores over all noisy condition are also given in this figure. By comparing the average values of Fig. 10 with the average values of *PESQ* criterion in Fig. 9, the general consistency of the subjective and objective evaluations can be conducted.

As it is shown in Fig. 10, the proposed methods have resulted in higher average of MOS scores than the other methods. The higher performance of the proposed methods

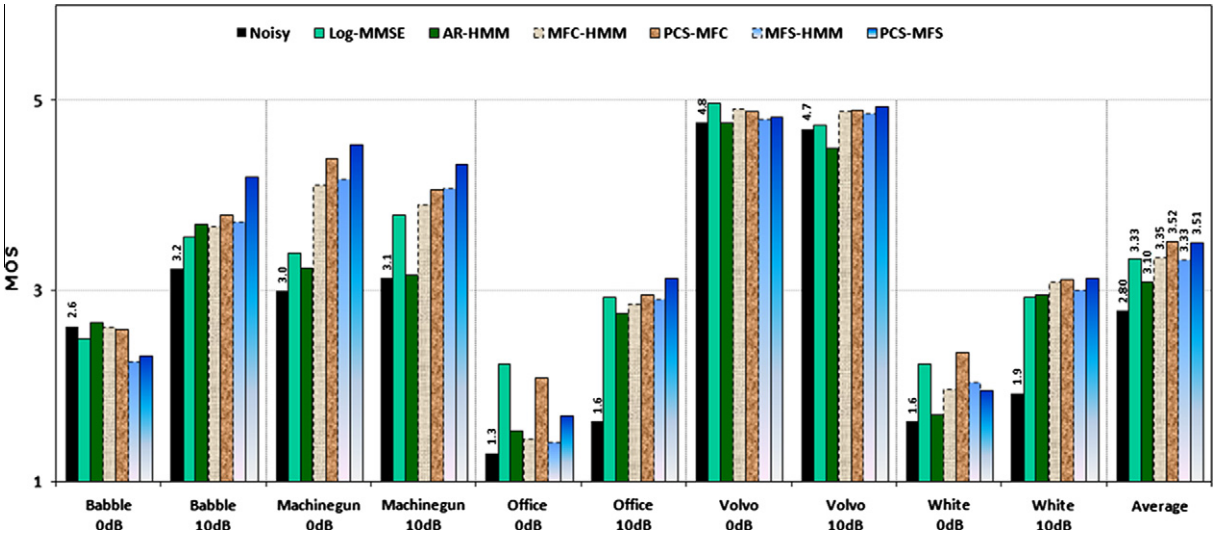


Fig. 10. MOS scores for selected reference methods and proposed methods in the presence of different noises at 0 and 10 dB. Last bars demonstrated the average over all noise types at all SNR levels.

is consistently observed in 10 dB condition for all noise types, however, the observation is different in 0 dB. In most experiments of 0 dB evaluations, the performances of MFC domain methods are higher than the MFS domain methods; this is consistent with the objective evaluation results given in the previous sections. This is probably due to the fact that speech spectrum is highly masked by noise in lower SNRs which has more corruption effect on the HMM modeling for the MFS features. Also, the MOS scores of the proposed methods in 0 dB conditions for babble and office noise types are lower than AR-HMM and Log-MMSE, respectively. As it can be observed from the figure, the MOS scores for Volvo noise are close to the maximum for all methods and even for the noisy signal in both SNR levels. It is because this noise is a low frequency noise the power of which is concentrated in frequencies lower than 500 Hz. Although this noise type affects the performance of the speech processing systems, it is not annoying for a human listener.

5. Summary and conclusions

In this paper, the HMM-based speech enhancement in the Mel-frequency domain was studied and evaluated particularly for the MFS and MFC coefficients. The problem of lossy inversion of the model parameters from the MFC/MFS domains to the spectral domain was addressed and PCS modeling was proposed to reduce this misrepresentation while utilizing the benefits of robust modeling in the Mel-frequency domain. Comprehensive evaluations and comparisons of the proposed methods and several established speech enhancement methods were performed with five noise types and three SNR levels. The proposed methods revealed to be superior to the reference methods at most testing conditions and resulted in significant improvements for non-stationary noise sources. The experimental results indicated that the proposed PCS modeling improved speech enhancement performance. These results also showed that the MFS coefficients performs better enhancement than the MFC ones. Finally, the results demonstrated that the proposed methods reduce the annoying non-stationary musical noise associated with the enhanced speech.

Acknowledgement

This research was partially supported by the Iranian Telecommunication Research Center (ITRC).

References

Arakawa, T., Tsujikawa, M., Isotani, R., 2006. Model-based Wiener filter for noise robust speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and, Signal Processing (ICASSP'06), pp. I-537–I540.

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'79), Washington DC, USA, pp. 208–211.

Chen, B., Loizou, P.C., 2007. A Laplacian-based MMSE estimator for speech enhancement. Speech Comm. 49 (2), 134–143.

Ephraim, Y., 1992a. A Bayesian estimation approach for speech enhancement using Hidden Markov models. IEEE Trans. Signal Process. 40 (4), 725–735.

Ephraim, Y., 1992b. Gain-adapted hidden Markov models for recognition of clean and noisy speech. IEEE Trans. Signal Process. 40 (6), 1303–1316.

Ephraim, Y., 1992c. Statistical model-based speech enhancement systems. Proc. IEEE 80 (10), 1526–1555.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'84), San Diego, California, USA, pp. 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'85), Florida, USA, pp. 443–445.

Ephraim, Y., Malah, D., Juang, B.H., 1989. On the application of hidden Markov models for enhancing noisy speech. IEEE Trans. Acoust. Speech Signal Process. 37 (12), 1846–1856.

Gales, M.J.F., 1995. Model-based Techniques for Noise Robust Speech Recognition, Ph.D. Thesis. University of Cambridge, September.

Griffin, D.W., Lim, J.S., 1984. Signal estimation from modified short-time fourier transform. IEEE Trans. Acoust. Speech Signal Process. ASSP-32 (2), 236–243 (April).

Hu, Y., Loizou, P., 2004. Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Trans. Acoust. Speech Signal Process. 12 (1), 59–67.

Imai, S., 1983. Cepstral analysis synthesis on the Mel frequency scale. In: Proceedings of IEEE International Conference on Acoustics, Speech, and, Signal Processing (ICASSP'83), pp. 93–96.

Lim, J., Oppenheim, A., 1978. All-pole modeling of degraded speech. IEEE Trans. Acoustics Speech Signal Process. 26 (3), 197–210.

Logan, B.T., 1998. Adaptive Model-Based Speech Enhancement, Ph.D. Thesis. Engineering Department, Cambridge University.

Loizou, P.C., 2007. Speech Enhancement: Theory and Practice. CRC Press, Boca Raton, FL.

Martin, R., 2005. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. IEEE Trans. Speech Audio Process. 13 (5 Part 2), 845–856.

Perceptual Evaluation of Speech Quality (PESQ), 2001. An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs, ITU-T Recommendation P.862, February.

Porter, J., Boll, S., 1984. Optimal estimators for spectral restoration of noisy speech. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'84), California, USA, pp. 53–56.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.

Sameti, H., 1994. Model-Based Approaches to Speech Enhancement: Stationary-State and Nonstationary-State HMMs, Ph.D. Thesis. Electrical Engineering, University of Waterloo, Waterloo, Canada.

Sameti, H., Sheikhzadeh, H., Deng, L., Brennan, R.L., 1998. HMM-based strategies for enhancement of speech signals embedded in nonstationary noise. IEEE Trans. Speech Audio Process. 6 (5), 445.

Sasou, A., Asano, F., Nakamura, S., Tanaka, K., 2006. HMM-based noise-robust feature compensation. Speech Comm. 48 (9), 1100–1111.

Segura, J.C., de la Torre, A., Benitez, M.C., Peinado, A.M., 2001. Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II database and tasks. In: Proceedings of EuroSpeech'01, vol. I. pp. 221–224.

Srinivasan, S., Samuelsson, J., Kleijn, W.B., 2007. Codebook-based Bayesian speech enhancement for nonstationary environments. IEEE Trans. Audio Speech Lang. Process. 15 (2), 441–452.

Stouten, V., Van Hamme, H., Wambacq, P., 2006. Model-based feature enhancement with uncertainty decoding for noise robust ASR. Speech Comm. 48 (11), 1502–1514.

Tokuda, K., Kobayashi, T., Imai, S., 1995. Speech parameter generation from HMM using dynamic features. In: Proceedings of IEEE International Conference on Acoustics, Speech, and, Signal Processing (ICASSP'95), pp. 660–663.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In: Proceedings of IEEE International Conference on Acoustics, Speech, and, Signal Processing (ICASSP'00), pp. 1315–1318.

Veisi, H., Sameti, H., 2011. The integration of principal component analysis and cepstral mean subtraction in parallel model combination for robust speech recognition. Digital Signal Process. 21 (1), 36–53.

Wolfe, P.J., Godsill, S.J., 2003. Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. EURASIP J. Appl. Signal Process. 2003 (10), 1043–1051.

You, C.H., Koh, S.N., Rahardja, S., 2003. Adaptive B-Order MMSE estimation for speech enhancement. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03), Hong Kong, China, pp. 852–855.

Yu, D., Deng, L., Droppo, J., Wu, J., Gong, Y., Acero, A., 2008. Robust speech recognition using a cepstral minimum-mean-square-error-motivated noise suppressor. IEEE Trans. Audio Speech Lang. Process. 16 (5), 1061–1070.

Zhao, D.Y., Kleijn, W.B., 2007. HMM-based gain modeling for enhancement of speech in noise. IEEE Trans. Audio Speech Lang Process. 15 (3), 882–892.