A novel NMF-HMM speech enhancement algorithm based on Poisson mixture model

Xiang, Yang; Shi, Liming; Lisby Højvang, Jesper ; Højfeldt Rasmussen, Morten ; Christensen, Mads Græsbøll

Published in:
ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

DOI (link to publication from Publisher):
10.1109/ICASSP39728.2021.9414620

Creative Commons License
CC BY-NC-ND 4.0

Publication date:
2021

Document Version
Accepted author manuscript, peer reviewed version

Citation for published version (APA):
Xiang, Y., Shi, L., Lisby Højvang, J., Højfeldt Rasmussen, M., & Christensen, M. G. (2021). A novel NMF-HMM speech enhancement algorithm based on Poisson mixture model. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 721-725). Article 9414620 IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414620

A NOVEL NMF-HMM SPEECH ENHANCEMENT ALGORITHM BASED ON POISSON MIXTURE MODEL

Yang Xiang[⋆†]*, Liming Shi*[⋆]*, Jesper Lisby Højvang*[†]*, Morten Højfeldt Rasmussen*[†]*, and*
Mads Græsbøll Christensen[⋆]

[⋆] Audio Analysis Lab, CREATE, Aalborg University, Aalbory, Denmark {yaxi,ls,mgc}@create.aau.dk
[†] Capturi A/S, Aarhus, Denmark {jlh,mhr}@capturi.com

ABSTRACT

In this paper, we propose a novel non-negative matrix factorization (NMF) and hidden Markov model (NMF-HMM) based speech enhancement algorithm, which employs a Poisson mixture model (PMM). Compared to the previously proposed NMF-HMM method, the new algorithm, termed PMM-NMF-HMM, uses the Poisson mixture distribution for the state conditional likelihood function for a HMM rather than the single Poisson distribution. This means that there are the more basis matrices that can be used to model the speech and noise signals, so more signal information can be captured by the resulting model. The proposed method is supervised and thus includes a training and an enhancement stage. It is shown that, in the training stage, the proposed method can be implemented efficiently using multiplicative update (MU) for the model parameters, much like the NMF-HMM algorithm. In the speech enhancement stage, which can be performed online, a novel PMM-NMF-HMM minimum mean-square error (MMSE) estimator is developed. The experimental results indicate that the PMM-NMF-HMM method can obtain higher short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) score than NMF-HMM. Additionally, the method also outperforms other state-of-the-art NMF-based supervised speech enhancement algorithms.

Index Terms— Poisson mixture model (PMM), speech enhancement, non-negative matrix factorization (NMF), hidden Markov model (HMM), minimum mean-square error (MMSE)

1. INTRODUCTION

In real-word environments,the quality and intelligibility of speech signal is often degraded due to the presence of background noise. To combat such noise, speech enhancement techniques have been developed. The main purpose of speech enhancement is to estimate the speech from the observed noisy speech while attenuating the background noise to improve the quality and intelligibility of the observed signal [1]. Monaural speech enhancement provides a cost-effective strategy to address this problem by utilizing recordings from a single microphone, and by combining it with beamforming it can be extended to multiple microphones. Speech enhancement has a wide rang of important applications, which include as automatic speech recognition (ASR) [2], teleconferencing, hearing-aids, and mobile communication.

During the past decades, many different speech enhancement strategies have been proposed for environments with additive noise (e.g., [3]). These methods can be roughly divided into supervised and unsupervised approaches. For the unsupervised algorithms, the spectral subtraction algorithm [4] is perhaps the simplest strategy to estimate the speech. Furthermore, the minimum mean-square error (MMSE) spectral amplitude estimator [5], the signal subspace method of [6] and the optimally-modified log-spectral amplitude (OM-LSA) method [7] combined with IMCRA noise estimator [8] are all effective strategies to estimate the speech. However, these methods cannot always achieve satisfactory speech enhancement performance in non-stationary noise environment because of inaccurate estimation of noise. Therefore, the supervised speech enhancement method have been proposed like NMF [9]. Among the supervised speech enhancement algorithms, the codebook-driven auto-regressive (AR) model based method [10], the auto-regressive hidden Markov model (ARHMM) method [11] and non-negative matrix factorization (NMF) based methods [12] are noteworthy methods. These algorithms can make good use of prior information about both speech and noise, and, as a result, they can often achieve better speech enhancement performance than the unsupervised methods, particularly in non-stationary acoustic environments.

With the advances in computation power, increases in the availability of training data combined with advances in the theory and practice of neural networks [13], deep neural networks (DNNs) have become a feasible strategy for speech enhancement. In recent years, various network structures have been used for enhancement, such as feed-forward multilayer perceptron [14], fully convolutional neural network [15], deep recurrent neural networks [16], and generative adversarial networks [17]. These networks can be used to predict the different targets like the speech spectrum [18], ideal ratio mask [19] and time domain waveform [20]. However, the computational complexity, model size and power consumption of these methods may be problematic for some application.

As mentioned above, NMF is an effective speech enhancement method. In general, NMF can be combined with other models to achieve the better speech enhancement performance. For instance, the combination of NMF and DNN can help NMF better model the speech and noise characteristics [21] and improve the generalization ability of the method [22]. Moreover, the NMF can be also combined with HMM [23], which can capture the temporal information of both speech and noise. As a consequence, such methods can often outperform the traditional NMF-based speech enhancement methods [12].

In our previous work [24], we proposed a NMF-HMM-based speech enhancement algorithm. This method applies a single Poisson distribution as the likelihood function for the HMM, which cannot effectively model the speech and noise due to their complex behavior. To address this problem we propose the Poisson Mixture Model-based NMF-HMM (PMM-NMF-HMM) speech enhancement algorithm, which is a more sophisticated statistical model

This work was partly supported by Innovation Fund Denmark (Grant No.9065-00046).

capable of capturing more complex behavior, similarly to Gaussian mixture models [25]. This model makes it possible to better describe the speech and noise because these may be governed by multiple underlying causes, each being responsible for one particular mixture component in the distribution. If such causes are identified, then the PMM-NMF-HMM can be decomposed into a set of cause-dependent or context-dependent component distributions [25]. As a result, the performance can, arguably, be improved by exploiting this. Furthermore, like the NMF-HMM-based speech enhancement algorithm, the proposed method can be implemented using multiplicative updates (MU) of the parameters. For performing the enhancement given the trained speech and noise models, we propose an PMM-NMF-HMM-based MMSE estimator, which can be implemented using online parameter updates suitable for parallel computations. Moreover, compared to typical DNN-based method [14], the proposed method uses a small model size with few degrees of freedom.

2. SIGNAL MODEL

In this section, we will briefly introduce the signal model that the proposed method is based. In an acoustic environment with additive noise, the observed signal model can be written as

$$y(l) = s(l) + d(l), \tag{1}$$

where $y(l)$, $s(l)$ and $d(l)$ represent the observed, speech and noise signals, respectively, and l is the time index. The short-time Fourier transform (STFT) of $y(l)$ can be written as

$$Y(f,n) = S(f,n) + D(f,n), \tag{2}$$

where $Y(f,n)$, $S(f,n)$, and $D(f,n)$ denotes the frequency spectrums of $y(l)$, $s(l)$, and $d(l)$, respectively. The f is the frequency bin index and the n is the time frame index. Collecting F frequency bins and N time frames, the magnitude spectrum matrices can be defined as $\mathbf{Y}_N$, $\mathbf{S}_N$ and $\mathbf{D}_N$, where $\mathbf{Y}_N = [\mathbf{y}_1, \cdots, \mathbf{y}_n, \cdots, \mathbf{y}_N]$ and $\mathbf{y}_n = [|Y(1,n)|, \cdots, |Y(f,n)|, \cdots, |Y(F,n)|]^T$, $\mathbf{s}_n$ and $\mathbf{d}_n$ are defined similarly to $\mathbf{y}_n$. And $\mathbf{S}_N$ and $\mathbf{D}_N$ are defined similarly to $\mathbf{Y}_N$. Additionally, the proposed method is based on the approximation $\mathbf{Y}_N \approx \mathbf{S}_N + \mathbf{D}_N$. The overbar $(\bar{\cdot})$ and double dots $(\ddot{\cdot})$ are used to represent the speech and the noise, respectively. The signal models for the speech and the noise signal are the same, so we will in what follows only shown them for the speech signal. Applying the conditional independence property of the standard HMM, the likelihood function for the speech can be expressed as follows:

$$p(\mathbf{S}_N;\mathbf{\Phi}) = \sum_{\overline{\mathbf{x}}_\mathbf{N}} \prod_{n=1}^{N} p(\mathbf{s}_n|\overline{x}_n)p(\overline{x}_n|\overline{x}_{n-1}), \tag{3}$$

where $\overline{\mathbf{x}}_\mathbf{N} = [\overline{x}_1, \cdots, \overline{x}_n, \cdots, \overline{x}_N]^T$ is a collection of states, $\overline{x}_n \in \{1, 2, \cdots, \overline{J}\}$ represents the state at the n^{th} frame and $\overline{J}$ denotes the total number of states. $p(\overline{x}_n|\overline{x}_{n-1})$ is the state transition probability from state $\overline{x}_{n-1}$ to $\overline{x}_n$ with $p(\overline{x}_1|\overline{x}_0)$ being the initial state probability. $p(\mathbf{s}_n|\overline{x}_n)$ is the state-conditioned likelihood function, $\mathbf{\Phi}$ is a collection of modeling parameters. In this work, we propose to apply PMM-NMF-HMM to estimate the $p(\mathbf{s}_n|\overline{x}_n)$, which can be written as

$$p(\mathbf{s}_n|\overline{x}_n) = \int p(\mathbf{s}_n|\overline{z}_n)p(\overline{z}_n|\overline{x}_n)\, d\overline{z}_n, \tag{4}$$

$$p(\overline{z}_n|\overline{x}_n) = \prod_{j=1}^{\overline{J}} \prod_{t=1}^{\overline{T}} \overline{P}_{j,t}^{\,l(\overline{x}_n=j,\overline{z}_n=t)}, \tag{5}$$

where $\overline{z}_n \in \{1, 2, \cdots, \overline{T}\}$ denotes the mixture state and $\overline{T}$ is the total number of mixture states. Additionally, we define $\overline{\mathbf{z}}_\mathbf{N} = [\overline{z}_1, \cdots, \overline{z}_n, \cdots, \overline{z}_N]^T$, which is a collection of mixture states. The $\overline{P}_{j,t}$ is the mixture weight and there is $\sum_{t=1}^{\overline{T}} \overline{P}_{j,t} = 1 (1 \le j \le \overline{J})$. The $l(\cdot)$ denotes an indicator function, which is 1 when the logical expression in the parentheses is true and 0 otherwise. In [26] it was demonstrated that the Kullback-Leibler (KL) divergence-based NMF can be derived from the following hierarchical statistical model:

$$\mathbf{S}_N = \sum_{k=1}^{\overline{K}} \overline{\mathbf{C}}(k), \tag{6}$$

$$\overline{c}_{f,n}(k) \sim \mathcal{PO}(\overline{c}_{f,n}(k); \overline{W}_{f,k}\overline{H}_{k,n}), \tag{7}$$

where $\mathcal{PO}(x;\lambda) = \lambda^x e^{-\lambda}/\Gamma(x+1)$ is the Poisson distribution, $\Gamma(x+1) = x!$ denotes the gamma function for positive integer x, $\overline{K}$ denotes the number of basis vectors, $\overline{\mathbf{C}}(k)$ is the latent matrix and $\overline{c}_{f,n}(k)$ denotes the element of $\overline{\mathbf{C}}(k)$ in the f^{th} row and n^{th} column. $\overline{W}_{f,k}$ and $\overline{H}_{k,n}$ correspond to the elements of the basis and activation matrices for the NMF. Based on the (6) and (7), we propose to apply the following hierarchical model to estimate $p(\mathbf{s}_n|\overline{z}_n)$,

$$\mathbf{s}_n = \sum_{k=1}^{\overline{K}} \overline{\mathbf{c}}_n(k), \tag{8}$$

$$p(\overline{\mathbf{c}}_n(k)|\overline{z}_n, \overline{x}_n) = \prod_{j,t,k,f} \{\mathcal{PO}(\overline{c}_{t,f,n}(k); \overline{W}_{f,k}^{\overline{x}_n,\overline{z}_n} \overline{H}_{k,n}^{\overline{x}_n,\overline{z}_n})\}^{l(\overline{x}_n=j,\overline{z_n}=t)}, \tag{9}$$

$$p(\mathbf{s}_n|\overline{\mathbf{c}}_n) = \delta(\mathbf{s}_n - \sum_k \overline{\mathbf{c}}_n(k)), \tag{10}$$

where $\overline{W}_{f,k}^{\overline{x}_n,\overline{z}_n}$ and $\overline{H}_{k,n}^{\overline{x}_n,\overline{z}_n}$ correspond to the elements of the basis and activation matrices and $\overline{\mathbf{c}}_n(k)$ contains the hidden variables, writing $\overline{\mathbf{c}}_n = [\overline{\mathbf{c}}_n(1)^T, \overline{\mathbf{c}}_n(2)^T, \cdots, \overline{\mathbf{c}}_n(\overline{K})^T]^T$ and integrating $\overline{\mathbf{c}}_n$ out, we obtain

$$\begin{aligned} p(\mathbf{s}_n|\overline{z}_n) &= \int p(\mathbf{s}_n|\overline{\mathbf{c}}_n)p(\overline{\mathbf{c}}_n|\overline{z}_n)\, d\overline{\mathbf{c}}_n \\ &= \prod_{j,t,f} \{\mathcal{PO}(|S(f,n)|; \sum_{k=1}^{\overline{K}} \overline{W}_{f,k}^{\overline{x}_n,\overline{z}_n} \overline{H}_{k,n}^{\overline{x}_n,\overline{z}_n})\}^{l(\overline{x}_n=j,\overline{z}_n=t)}. \end{aligned} \tag{11}$$

Finally, combining (4) and (5), at jth state, the (11) can be written as

$$\begin{aligned} &p(\mathbf{s}_n|\overline{x}_n = j) = \\ &\sum_{t=1}^{\overline{T}} \overline{P}_{j,t} \prod_{f=1}^{F} \mathcal{PO}(|S(f,n)|; \sum_{k=1}^{\overline{K}} \overline{W}_{f,k}^{\overline{x}_n,\overline{z}_n} \overline{H}_{k,n}^{\overline{x}_n,\overline{z}_n}) \end{aligned} \tag{12}$$

Moreover, we have that

$$\begin{aligned} &p(\mathbf{s}_n|\overline{x}_n, \overline{z}_n) = \\ &\mathcal{PO}(|S(f,n)|; \sum_{k=1}^{\overline{K}} \overline{W}_{f,k}^{\overline{x}_n,\overline{z}_n} \overline{H}_{k,n}^{\overline{x}_n,\overline{z}_n}). \end{aligned} \tag{13}$$

We collect the unknown parameters $\{\overline{W}_{f,k}^{\overline{x}_n,\overline{z}_n}\}$ and $\{\overline{H}_{k,n}^{\overline{x}_n,\overline{z}_n}\}$ in matrices $\{\overline{\mathbf{W}}^{\mathbf{j,t}}\}$ and $\{\overline{\mathbf{H}}^{\mathbf{j,t}}\}$. To summarize, there are five parameters to be estimated in our proposed clean speech model. They are the initial state probability matrix $\overline{\boldsymbol{\pi}}$, state transition probability matrix $\overline{\mathbf{A}}$, basis matrix $\overline{\mathbf{W}}^{\mathbf{j,t}}$, activation matrix $\overline{\mathbf{H}}^{\mathbf{j,t}}$ and mixture weight matrix $\overline{\mathbf{P}}$. The activation matrix $\overline{\mathbf{H}}^{\mathbf{j,t}}$ is estimated in the online speech enhancement stage while the other parameters are obtained in the offline training stage. Additionally, the $\overline{K}$ and $\overline{T}$ can be predetermined. For the observed signal, the initial state and transition probabilities matrix can be expressed as $\overline{\boldsymbol{\pi}} \otimes \ddot{\boldsymbol{\pi}}$ and $\overline{\mathbf{A}} \otimes \ddot{\mathbf{A}}$, where the $\otimes$ denotes the Kronecker product. Thus, the conditional likelihood function can be written as

$$\begin{aligned} p(\mathbf{y}_n|\overline{x}_n, \ddot{x}_n) = \sum_{t=1}^{\overline{T}} \sum_{t=1}^{\ddot{T}} \overline{P}_{j,t} \ddot{P}_{j,t} \prod_{f=1}^{F} \mathcal{PO}(|Y(f,n)|; \\ \sum_{k=1}^{\overline{K}} \overline{W}_{f,k}^{\overline{x}_n,\overline{z}_n} \overline{H}_{k,n}^{\overline{x}_n,\overline{z}_n} + \sum_{k=1}^{\ddot{K}} \ddot{W}_{f,k}^{\ddot{x}_n,\ddot{z}_n} \ddot{H}_{k,n}^{\ddot{x}_n,\ddot{z}_n}) \end{aligned} \tag{14}$$

$$\begin{aligned} p(\mathbf{y}_n|\overline{z}_n, \ddot{z}_n, \overline{x}_n, \ddot{x}_n) = \prod_{f=1}^{F} \mathcal{PO}(|Y(f,n)|; \\ \sum_{k=1}^{\overline{K}} \overline{W}_{f,k}^{\overline{x}_n,\overline{z}_n} \overline{H}_{k,n}^{\overline{x}_n,\overline{z}_n} + \sum_{k=1}^{\ddot{K}} \ddot{W}_{f,k}^{\ddot{x}_n,\ddot{z}_n} \ddot{H}_{k,n}^{\ddot{x}_n,\ddot{z}_n}). \end{aligned} \tag{15}$$

3. OFFLINE PARAMETER ESTIMATION

As mentioned above, the algorithm can be divided into two stages. In the offline training stage, the parameters of speech and noise signal model are estimated by using the speech and noise database, respectively. First, we define the complete data set $(\mathbf{S}_N, \overline{\mathbf{x}}_N, \overline{\mathbf{z}}_N, \overline{\mathbf{C}}_N)$, where $\overline{\mathbf{C}}_N = [\overline{\mathbf{c}}_1, \overline{\mathbf{c}}_2, \cdots, \overline{\mathbf{c}}_N]$. Based on the (3) and derivation in Section 2, using the conditional independence property, the complete data likelihood function can be written as

$$\begin{aligned} & p(\mathbf{S}_N, \overline{\mathbf{x}}_N, \overline{\mathbf{z}}_N, \overline{\mathbf{C}}_N) \\ & = \left(\prod_{n=1}^{N} p(\mathbf{s}_n|\overline{\mathbf{c}}_n)\right) \left(p(x_1) \prod_{n=2}^{N} p(\overline{\mathbf{x}}_n|\overline{\mathbf{x}}_{n-1})\right) \\ & \left(\prod_{n=1}^{N} p(\overline{\mathbf{z}}_n|\overline{\mathbf{x}}_n)\right) \left(\prod_{n=1}^{N} p(\overline{\mathbf{c}}_n|\overline{\mathbf{x}}_n, \overline{\mathbf{z}}_n)\right). \end{aligned} \tag{16}$$

Using Expectation–Maximization (EM) algorithm [27], the model parameters can be estimated. For simplicity, we here omit the derivation process. It can be shown that the parameter updates can be written as follows:

$$\overline{\pi}_j = \frac{q(\overline{x}_1 = j)}{\sum_{o=1}^{\overline{J}} q(\overline{x}_1 = o)}, \tag{17}$$

$$\overline{A}_{o,j} = \frac{\sum_{n=2}^{N} q(\overline{x}_n = j, \overline{x}_{n-1} = o)}{\sum_{j=1}^{\overline{J}} \sum_{n=2}^{N} q(\overline{x}_n = j, \overline{x}_{n-1} = o)}, \tag{18}$$

where $1 \leq o, j \leq \overline{J}$. The quantities $q(\overline{x}_n)$ and $q(\overline{x}_n, \overline{x}_{n-1})$ correspond to the posterior state probability and the joint posterior probability, which can be calculated by forward-backward algorithm [24] that combines the (12). The $\overline{\pi}_j$ and $\overline{A}_{o,j}$ is the elements of $\overline{\mathbf{A}}$ and $\overline{\boldsymbol{\pi}}$, respectively. The estimation of $\overline{\mathbf{A}}$ and $\overline{\boldsymbol{\pi}}$ is similar to the traditional HMM. In addition, we have the following updates:

$$\overline{\mathbf{W}}^{\mathbf{j,t}} \leftarrow \overline{\mathbf{W}}^{\mathbf{j,t}} \odot \frac{\frac{\mathbf{S_N}}{\overline{\mathbf{W}}^{\mathbf{j,t}} \overline{\mathbf{H}}^{\mathbf{j,t}}} \boldsymbol{\Lambda}(\mathbf{j},\mathbf{t}) (\overline{\mathbf{H}}^{\mathbf{j,t}})^T}{\mathbf{1} \boldsymbol{\Lambda}(\mathbf{j},\mathbf{t}) (\overline{\mathbf{H}}^{\mathbf{j,t}})^T}, \tag{19}$$

$$\overline{\mathbf{H}}^{\mathbf{j,t}} \leftarrow \overline{\mathbf{H}}^{\mathbf{j,t}} \odot \frac{(\overline{\mathbf{W}}^{\mathbf{j,t}})^T \frac{\mathbf{S_N}}{\overline{\mathbf{W}}^{\mathbf{j,t}} \overline{\mathbf{H}}^{\mathbf{j,t}}}}{(\overline{\mathbf{W}}^{\mathbf{j,t}})^T \mathbf{1}}, \tag{20}$$

where $\boldsymbol{\Lambda}(\mathbf{j},\mathbf{t}) = \mathrm{diag}(q(\overline{x}_1 = j, \overline{z}_1 = t), q(\overline{x}_2 = j, \overline{z}_2 = t), \cdots, q(\overline{x}_N = j, \overline{z}_N = t))$. The $q(\overline{x}_n = j, \overline{z}_n = t)$ is the posterior probability when $(\overline{x}_n = j, \overline{z}_n = t)$. Once again, this calculation can be performed using the forward-backward algorithm which uses (13). Furthermore, this update is in the form of an multiplicative update, which means that the offline training can be performed efficiently. Moreover, we have

$$\overline{P}_{j,t} = \frac{\sum_{n=1}^{N} q(\overline{x}_n = j, \overline{z}_n = t)}{\sum_{n=1}^{N} \sum_{t=1}^{\overline{T}} q(\overline{x}_n = j, \overline{z}_n = t)} \tag{21}$$

This mixture weight $\overline{P}_{j,t}$ determines the importance of each latent cause that is modeled by single Poisson distribution for the whole speech signal.

4. ONLINE SPEECH ENHANCEMENT

In the online enhancement stage, we propose a novel MMSE estimator, which is based on the model produced by the PMM-NMF-HMM algorithm. The MMSE estimate of the speech signal from the noisy observation is

$$\hat{\mathbf{s}}_n = \mathbb{E}_{\mathbf{s}_n|\mathbf{Y}_n}(\mathbf{s}_n) = \int \mathbf{s}_n p(\mathbf{s}_n|\mathbf{Y}_n)\, d\mathbf{s}_n. \tag{22}$$

For simplicity, we omit the specific details of this derivation. The enhanced speech can be written as $\hat{\mathbf{s}}_n = \mathbf{y}_n \odot \mathbf{g}_n$ where $\mathbf{g}_n$ can be viewed as a spectral gain vector with

$$\mathbf{g}_n = \sum_{\overline{x}_n, \ddot{x}_n} \omega_{\overline{x}_n, \ddot{x}_n} \left(\sum_{\overline{z}_n, \ddot{z}_n} \overline{P}_{j,t} \ddot{P}_{j,t} \mathbf{p}_n(\overline{x}_n, \ddot{x}_n, \overline{z}_n, \ddot{z}_n) \right), \tag{23}$$

where the weight $0 \leq \omega_{\overline{x}_n, \ddot{x}_n} \leq 1$ can be written as

$$\omega_{\overline{x}_n, \ddot{x}_n} = \frac{p(\mathbf{y}_n|\overline{x}_n, \ddot{x}_n) p(\overline{x}_n, \ddot{x}_n|\mathbf{Y}_{n-1})}{\sum_{\overline{x}_n, \ddot{x}_n} p(\mathbf{y}_n|\overline{x}_n, \ddot{x}_n) p(\overline{x}_n, \ddot{x}_n|\mathbf{Y}_{n-1})}. \tag{24}$$

The calculation of $p(\mathbf{y}_n|\overline{x}_n, \ddot{x}_n)$ can be conducted using (14), and

$$\begin{aligned} & p(\overline{x}_n, \ddot{x}_n|\mathbf{Y}_{n-1}) \\ & = \sum_{\overline{x}_{n-1}, \ddot{x}_{n-1}} p(\overline{x}_n, \ddot{x}_n|\overline{x}_{n-1}, \ddot{x}_{n-1}, \mathbf{Y}_{n-1}) p(\overline{x}_{n-1}, \ddot{x}_{n-1}|\mathbf{Y}_{n-1}) \\ & = \sum_{\overline{x}_{n-1}, \ddot{x}_{n-1}} p(\overline{x}_n, \ddot{x}_n|\overline{x}_{n-1}, \ddot{x}_{n-1}) p(\overline{x}_{n-1}, \ddot{x}_{n-1}|\mathbf{Y}_{n-1}), \end{aligned} \tag{25}$$

In (25), the first term can be calculated by the transition probabilities matrix of observed signal and the second term is the forward probability which can be calculated by a forward algorithm [24]. In (23), $\mathbf{p}_n(\overline{x}_n, \ddot{x}_n, \overline{z}_n, \ddot{z}_n) = [p_{1,n}(\overline{x}_n, \ddot{x}_n, \overline{z}_n, \ddot{z}_n), p_{2,n}(\overline{x}_n, \ddot{x}_n, \overline{z}_n, \ddot{z}_n), \cdots, p_{F,n}(\overline{x}_n, \ddot{x}_n, \overline{z}_n, \ddot{z}_n)]^T$, where

$$\begin{aligned} & p_{f,n}(\overline{x}_n, \ddot{x}_n, \overline{z}_n, \ddot{z}_n) = \\ & \frac{\sum_{k=1}^{\overline{K}} \overline{W}_{f,k}^{\overline{x}_n,\overline{z}_n} \overline{H}_{k,n}^{\overline{x}_n,\overline{z}_n}}{\sum_{k=1}^{\overline{K}} \overline{W}_{f,k}^{\overline{x}_n,\overline{z}_n} \overline{H}_{k,n}^{\overline{x}_n,\overline{z}_n} + \sum_{k=1}^{\ddot{K}} \ddot{W}_{f,k}^{\ddot{x}_n,\ddot{z}_n} \ddot{H}_{k,n}^{\ddot{x}_n,\ddot{z}_n}}, \end{aligned} \tag{26}$$

Comparing the PMM-NMF-HMM-based MMSE estimator with our previous proposed NMF-HMM-based MMSE estimator [24], we can

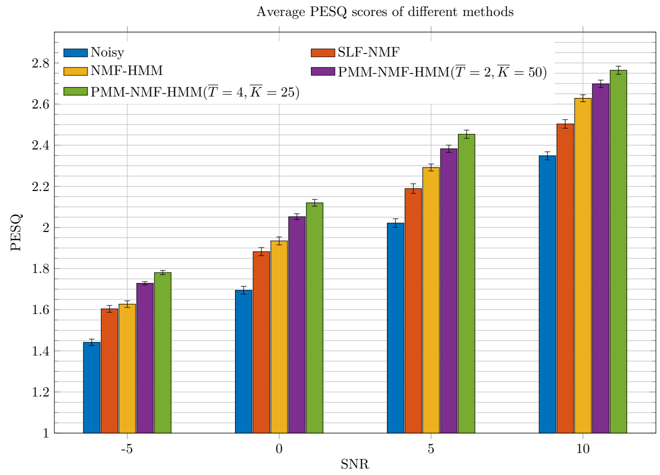

Fig. 1. Average PESQ scores of different algorithms using six types of noise under four different SNRs.

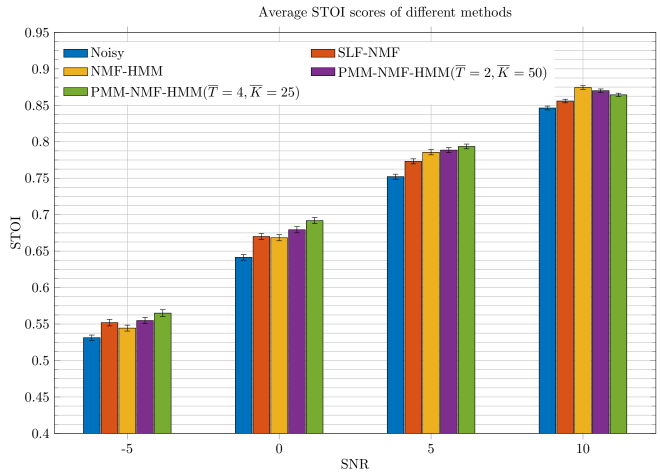

Fig. 2. Average STOI scores of different algorithms using six types of noise under four different SNRs.

find that there are more than one NMF basis matrices at each HMM hidden state, which means that our algorithm can model more underlying possible causes in the observed signal, so the enhancement performance can likely be improved based on this better model. Furthermore, we also remark that parallel computing can be applied to conduct the online estimation of active matrix $\overline{\overline{\mathbf{H}}}^{\mathbf{j,t}}$ to effectively reduce the time consumption.

5. EXPERIMENTAL RESULT AND ANALYSIS

In this section, the performance of proposed algorithm was evaluated and compared with state-of-the-art NMF-based speech enhancement algorithms. All the experiments were conducted on the TIMIT [28] and NOISEX-92 [29] databases. In the training stage, all 4620 utterances from the training set of the TIMIT database were used to train the speech PMM-NMF-HMM model. Meanwhile, parts of the Babble, F16, Factory and White noise from the NOISEX-92 database were used to train the noise model. In the test stage, 200 utterances were randomly chosen from the test set of the TIMIT database. After that, the chosen 200 utterances were added to six types of noise at four different SNR levels (i.e., -5, 0, 5, and 10 dB). There were two types of noise (destroyerengine and destroyerops) that were not included in the training database to test the generalization ability of the noise model. It must be stressed that for all noise types, disjoint training and test data was used.

To evaluate the performance of the proposed method, we compare to two state-of-the-art methods, namely the NMF-HMM [24] and the variable span linear filters [6] (SLF-NMF) combined with parametric NMF [10] for estimating the noise and speech statsitics.

In the experiments, all the signal waveform was down-sampled to 16 kHz. The frame length was set to 1024 samples with a frame shift of 512 samples . The size of STFT was 1024 points with a Hanning window. Furthermore, the maximum number of iterations was set to 30 in the training stage and 15 in the online speech enhancement stage for these NMF-based methods. In addition, the PESQ [30], ranging from -0.5 to 4.5, was used to evaluate the enhanced speech quality. The STOI [31], ranging from 0 to 1, was used to measure the enhanced speech intelligibility.

For the NMF parameter setting, to better compare the performance of PMM-NMF-HMM and NMF-HMM, we ensure that there are the same total number of basis vector for the two models. For the NMF-HMM, there is no the mixture weight (the NMF-HMM can be seen as a special case of PMM-NMF-HMM when $\overline{T}=1$ and $\ddot{T}=1$), so we only need to set $\overline{J}=10, \overline{K}=100, \ddot{J}=2$ and $\ddot{K}=70$. For the PMM-NMF-HMM, we have $\overline{J}=10, \ddot{J}=2, \ddot{K}=70, \ddot{T}=1$. We investigate the two different $\overline{T}$. When $\overline{T}$ is set to 2 and 4, the $\overline{K}$ corresponds to 50 and 25, which ensures that there is the same number of total basis vector. For the SLF-NMF, we utilize the maximum SNR filter and the codebook size of speech and noise is set to 64 and 8, respectively. Figure 1 indicates the average PESQ scores with 95% confidence interval of these algorithms. The NMF-HMM-based methods always achieve higher PESQ scores than SLF-NMF for all four SNRs. Additionally, with increased total number of mixture state $\overline{T}$, PMM-NMF-HMM achieve the better performance. This indicates that PMM-NMF-HMM may effectively better model multiple underlying causes in speech and noise when improving the speech quality. Figure 2 shows the average STOI scores with 95% confidence interval of the methods. We can see that the PMM-NMF-HMM achieves better speech enhancement performance at low SNRs (-5, 0, 5dB) with increased numbers of mixture state $\overline{T}$. However, for high SNRs, more mixture states does not lead to a better performance.

6. CONCLUSION

In this work, we have proposed a novel PMM-NMF-HMM-based speech enhancement algorithm. The new method employs a PMM which was used to model the state-conditioned likelihood function for the HMM, whereby multiple underlying causes in the signals could be captured. More specifically, the resulting modal can be decomposed into the different sets of cause-dependent or context-dependent component distributions. Finally, as a result of the new and more sophisticated model, the speech can be estimated more accurately. To enhance the speech, we have proposed a novel MMSE estimator, which is also based on the model of the PMM-NMF-HMM method. This estimator can be implemented efficiently and is thus suitable for online speech enhancement. In general, the experimental results showed that the proposed PMM-NMF-HMM method outperforms the previously proposed NMF-HMM, though the STOI score was slightly lower than NMF-HMM at high SNR (10dB).

7. REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 4, pp. 745–777, 2014.

[3] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 873–902.

[4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

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

[6] J. R. Jensen, J. Benesty, and M. G. Christensen, "Noise reduction with optimal variable span linear filters," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 4, pp. 631–644, 2015.

[7] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, 2001.

[8] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[9] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Advances in Neural Inform. Process. Syst.*, 2001, pp. 556–562.

[10] M. S. Kavalekalam, J. K. Nielsen, L. Shi, M. G. Christensen, and J. Boldt, "Online parametric NMF for speech enhancement," in *Proc. European Signal Processing Conf.*, 2018, pp. 2320–2324.

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

[12] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," in *Int. Conf. Digital Signal Process.*, 2011, pp. 1–6.

[13] Y. Bengio *et al.*, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[14] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, 2013.

[15] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, 2016.

[16] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, 2015.

[17] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[18] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 136–140.

[19] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[20] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[21] T. T. Vu, B. Bigot, and E. S. Chng, "Combining non-negative matrix factorization and deep neural networks for speech enhancement and automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 499–503.

[22] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 101–105.

[23] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, 2013.

[24] Y. Xiang, L. Shi, J. L. Højvang, M. H. Rasmussen, and M. G. Christensen, "An nmf-hmm speech enhancement method based on kullback-leibler divergence," in *Proc. Interspeech*, 2020, pp. 2667–2671.

[25] D. Yu and L. Deng, *Automatic Speech Recognition*. Springer, 2016.

[26] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational intelligence and neuroscience*, vol. 2009, 2009.

[27] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, no. 1, pp. 1–8, 1972.

[28] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[29] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 2001, pp. 749–752.

[31] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.