An NMF-HMM Speech Enhancement Method based on Kullback-Leibler Divergence

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

Published in:
Interspeech

Publication date:
2020

An NMF-HMM Speech Enhancement Method based on Kullback-Leibler Divergence

Yang Xiang[1,2]*, Liming Shi*[1]*, Jesper Lisby Højvang*[2]*, Morten Højfeldt Rasmussen* [2]*, Mads Græsbøll Christensen*[1]

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

Abstract

In this paper, we present a novel supervised Non-negative Matrix Factorization (NMF) speech enhancement method, which is based on Hidden Markov Model (HMM) and Kullback-Leibler (KL) divergence (NMF-HMM). Our algorithm applies the HMM to capture the timing information, so the temporal dynamics of speech signal can be considered by comparing with the traditional NMF-based speech enhancement method. More specifically, the sum of Poisson, leading to the KL divergence measure, is used as the observation model for each state of HMM. This ensures that the parameter update rule of the proposed algorithm is identical to the multiplicative update rule, which is quick and efficient. In the training stage, this update rule is applied to train the NMF-HMM model. In the online enhancement stage, a novel minimum mean-square error (MMSE) estimator that combines the NMF-HMM is proposed to conduct speech enhancement. The performance of the proposed algorithm is evaluated by perceptual evaluation of speech quality (PESQ) and short-timeobjective intelligibility (STOI). The experimental results indicate that the STOI score of proposed strategy is able to outperform 7% than current state-of-the-art NMF-based speech enhancement methods.

Index Terms: speech enhancement, non-negative matrix factorization, hidden markov model, minimum mean-square error

1. Introduction

The aim of single-channel speech enhancement (SE) is to remove background noise from the noisy environment to improve quality and intelligibility of noisy speech. Nowadays, SE has achieved a wide range of applications in hearing aids, mobile communication, robust speech recognition (ASR) [1], teleconferencing and speech coding etc. Therefore, during the past decades, many different approaches have been proposed [2].

In an environment with additive noise, the spectral subtraction algorithm [3] is the simplest strategy to achieve SE, which subtracts the noise spectrum from the observed signal. Furthermore, some unsupervised algorithms like Wiener filtering [4], signal subspace algorithm [5], minimum mean-square error (MMSE) spectral amplitude estimator [6] and log-MMSE spectral amplitude estimator [7] are also the effective strategies to conduct the SE. However, these methods cannot always achieve satisfactory performance in the non-stationary noisy environment because they are usually based on some inaccurate assumptions and do not apply the prior information of clean speech and noise.

As a result, some supervised SE methods have been developed. These approaches usually consider to train a model and the model parameters are acquired by using the speech and noise signals. These methods include codebook-based algorithms [8], Hidden Markov Model (HMM)-based strategies [9] and Deep Neural Network (DNN)-based approaches [10–12] etc. These algorithms can make use of the prior information of clean speech and noise, so they can achieve better speech enhancement performance in practical noisy environments.

Non-negative Matrix Factorization (NMF)-based [13] [14] SE method can be also viewed as such a kind of supervised speech enhancement strategy. In paper [15], a mask-based NMF SE was proposed, which trained the basis matrix of clean speech and noise during offline stage. On the enhancement stage, the activation matrix could be acquired by combining the trained basis matrix and noisy signal. After that, the mask was estimated for the application of speech enhancement. In paper [16], an NMF-based denoising scheme was proposed. This method added a heuristic term to the cost function, so the NMF coefficient can be adjusted according to the long-term levels of signals. Smaragdis et al. [17] proposed a supervised and unsupervised NMF speech enhancement method. In [17], the noise basis matrix could be acquired by combining the HMM during the enhancement stage. Thus, this method could mitigate the problem of noise mismatch. Furthermore, a NMF-based source separation approach was proposed in paper [18], which considers the HMM.

Inspired by these previous studies, in this paper, we proposed a novel NMF-HMM speech enhancement algorithm, which applies the Kullback-Leibler (KL) divergence. Compared to most NMF-based methods [13] [14], our method can utilize the temporal dynamics of speech signals to conduct the speech enhancement, so the time information of speech signal can be considered. Moreover, we used the sum of Poisson distribution as the state conditioned likelihood for the HMM rather than the general Gaussian Mixture Model (GMM), because the sum of Poisson distribution leads to the KL divergence measure, which is a mainstream measure in NMF, and its parameter update rule is identical to the multiplicative update rule. This ensures the parameter update is fast and efficient. On the enhancement stage, a minimum mean-square error (MMSE) estimator was derived to conduct SE, which was based on the NMF and HMM. The benefit of this algorithm is that the update of activation matrix can be conducted by parallel computing, which reduces the computation time.

2. NMF-based Speech Enhancement with KL divergence

In this section, we will briefly review the NMF-based speech enhancement method with KL divergence. In this work, we only consider to achieve speech enhancement in the additive noisy

environment. Thus, the noisy signal model can be represented as following:

$$y(t) = s(t) + m(t), \tag{1}$$

where $y(t)$, $s(t)$ and $m(t)$ are the noisy speech, clean speech and noise, respectively. The t is the time index. The short time Fourier transform (STFT) of $y(t)$ can be written as

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

where $Y(f,n)$, $S(f,n)$ and $M(f,n)$ are the complex STFT parameters of $y(t)$, $s(t)$ and $m(t)$, respectively. The f denotes frequency bin index and the n is the time frame index. For the sake of simplicity, we omit the frequency bin index, so their magnitude can be rewritten as the vectors $\mathbf{Y}_n$, $\mathbf{S}_n$ and $\mathbf{M}_n$.

For the NMF analysis, the magnitude of a signal $\mathbf{V}$ can be represented as

$$\mathbf{V} \approx \mathbf{WH}, \tag{3}$$

where $\mathbf{W}$ denotes the basis matrix and $\mathbf{H}$ denotes the activation matrix. Based on KL divergence, $\mathbf{W}$ and $\mathbf{H}$ can be estimated using iterative multiplicative update rules [14]

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\mathbf{WH}}\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T}, \tag{4}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T\frac{\mathbf{V}}{\mathbf{WH}}}{\mathbf{W}^T\mathbf{1}}, \tag{5}$$

where the $\odot$ and all divisions are element-wise multiplication and division operations, respectively. The $\mathbf{1}$ is the matrix of ones with the same size of $\mathbf{V}$ and T is the matrix transpose. For the application of speech enhancement, the speech basis matrix $\overline{\mathbf{W}}$ and noise basis matrix $\ddot{\mathbf{W}}$ can be estimated from clean speech and noise during the training stage. On the enhancement stage, the noisy speech basis matrix can be acquired by $\mathbf{W} = [\overline{\mathbf{W}}, \ddot{\mathbf{W}}]$. Additionally, the activation matrix $\mathbf{H}$ of noisy speech can be estimated by (5). After obtaining $\mathbf{H}$ and $\mathbf{W}$, the speech enhancement an be conducted by various algorithms [15] [16] [17] [18].

Furthermore, the [19] proves that the NMF with the KL divergence can be also motivated from the following hierarchical probability model

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

$$c_{f,n} \sim \mathcal{PO}(c_{f,n}(k); W_{f,k}H_{k,n}), \tag{7}$$

where the Poisson distribution $\mathcal{PO}(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{\Gamma(x+1)}$, and $\Gamma(x+1) = x!$ is the Gamma function, K is the number of basis vectors, and $c_{f,n}$ is the latent variable of $\mathbf{C}(k)$ for Poisson distribution. Note, the $\mathbf{V}$ is assumed Poisson-distributed and integer-valued. In practice, the factorial is approximated by the Gamma function [19]. It has been shown that [19] the iterative update of the parameters $\mathbf{H}$ and $\mathbf{W}$ using Expectation–Maximization (EM) algorithm is identical to the multiplicative update rules (4) and (5).

3. NMF-HMM-based Speech Enhancement

In this section, the details of the proposed algorithm will be illustrated, which includes the proposed signal model, offline parameter learning and online speech enhancement.

3.1. HMM-based signal models with the KL divergence

In our proposed approach, we need to acquire the three different signal models. They are namely clean speech model, noise model and noisy speech model. They will be separately introduced in this part. We use the overbar and double dots to represent the clean speech and noise, respectively.

In this work, there is the same signal model for the clean speech and the noise signal, so we will illustrate them just using clean speech signal. In order to model clean speech $\mathbf{S}_n$, we propose to a novel NMF-HMM-based method. To acquire a HMM model, there are three parameters [20] to be estimated. They are initial state probability $\overline{\boldsymbol{\pi}}$, transition probability matrix $\overline{\mathbf{A}}$ and state conditioned likelihood function. In addition, there are total $\overline{J}$ hidden states for this model. Thus, based on (6), we propose to model $\mathbf{S_n}$ as

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

By applying the (7) and HMM [16], for the jth ($j = 1, 2, \cdots, \overline{J}$) state, we can be defined

$$p(\overline{\mathbf{c}}_n|\overline{x}_n) = \prod_{f=1}^{F}\prod_{k=1}^{\overline{K}} \mathcal{PO}(\overline{c}_{f,n}(k); \overline{W}_{f,k}^{\overline{x}_n}\overline{H}_{k,n}^{\overline{x}_n}), \tag{9}$$

where the $\overline{x}_n$ is the hidden state and $\overline{x}_n \in \{1, 2, \cdots, \overline{J}\}$. $\overline{K}$ is the number of basis of clean speech and F is the total number of frequency bins. $\overline{W}_{k,n}^{\overline{x}_n}$ and $\overline{H}_{k,n}^{\overline{x}_n}$ is corresponding to the elements of the basis and activation for clean speech. Thus, the conditioned likelihood function at the jth state can be finally written as

$$p(\mathbf{S}_n|\overline{x}_n) = \prod_{f=1}^{F} \mathcal{PO}(S(f,n); \sum_{k=1}^{\overline{K}} \overline{W}_{f,k}^{\overline{x}_n}\overline{H}_{k,n}^{\overline{x}_n}), \tag{10}$$

where we use the superposition property of Poisson random variable. From (9), it can be found that there are $\overline{J}$ basis matrices for speech modelling, instead of one basis matrix in the traditional NMF, which is able to effectively capture the temporal dynamics of speech signals. The benefits of choosing the sum of Poission distribution as the state conditioned likelihood function is that its parameters update rules using EM algorithm is identical to the multiplicative update rules leading to low computational complexity. In addition, it is based on non-negative data by comparing with traditional HMM.

To sum up, the proposed model includes four parameters ($\overline{\mathbf{A}}$, $\overline{\boldsymbol{\pi}}$, $\overline{\mathbf{W}}^{\overline{x}_n}$ and $\overline{\mathbf{H}}^{\overline{x}_n}$). The $\overline{\mathbf{H}}^{\overline{x}_n}$ can be estimated by online speech enhancement and the other three parameters can be obtained by offline learning.

Based on proposed clean speech, noise signal model and (2), the noisy speech model can be defined. We assume that there are $\ddot{J}$ hidden states for noise and the hidden state of noise is $\ddot{x}_n (\ddot{x}_n \in \{1, 2, \cdots, \ddot{J}\})$. The $\ddot{\boldsymbol{\pi}}$ and $\ddot{\mathbf{A}}$ represent the initial state probability and transition probability matrix of the noise. Thus, there are total $\overline{J} \times \ddot{J}$ hidden states for noisy speech. The initial state and transition probabilities matrix of noisy speech 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. Finally, the conditioned likelihood function of noisy speech can be written as

$$p(\mathbf{Y}_n|\overline{x}_n, \ddot{x}_n) = \prod_{f=1}^{F} \mathcal{PO}(Y(f,n); \sum_{k=1}^{\overline{K}} \overline{W}_{f,k}^{\overline{x}_n}\overline{H}_{k,n}^{\overline{x}_n} + \sum_{k=1}^{\ddot{K}} \ddot{W}_{f,k}^{\ddot{x}_n}\ddot{H}_{k,n}^{\ddot{x}_n}), \tag{11}$$

where $\ddot{K}$, $\ddot{W}_{k,n}$ and $\ddot{H}_{k,n}$ is the number of basis, elements of the basis matrices and activation for noise.

3.2. Offline NMF-HMM parameter learning

In offline training stage, the aim is to find the parameter set $\mathbf{\Phi}$ to maximize the likelihood function, which is based on the HMM and EM algorithm [20]. There is the similar process for the parameter learning of clean speech and noise, so we will use the clean speech as the example to illustrate this process. At first, we define the complete data set $(\mathbf{S_N}, \overline{\mathbf{X}}_\mathbf{N}, \overline{\mathbf{C}}_\mathbf{N})$, where $\mathbf{S_N} = [\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_N]$, $\overline{\mathbf{X}}_\mathbf{N} = [\overline{x}_1, \overline{x}_2, \cdots, \overline{x}_N]^T$ and $\overline{\mathbf{C}}_\mathbf{N} = [\overline{\mathbf{c}}_1, \overline{\mathbf{c}}_2, \cdots, \overline{\mathbf{c}}_N]$. Thus, the complete data likelihood can be written as

$$p(\mathbf{S_N}, \overline{\mathbf{X}}_\mathbf{N}, \overline{\mathbf{C}}_\mathbf{N}) = \prod_{n=1}^{N} p(\mathbf{S}_n|\overline{\mathbf{c}}_n)p(\overline{\mathbf{c}}_n|\overline{x}_n)p(\overline{x}_n|\overline{x}_{n-1}). \tag{12}$$

By applying the EM algorithm in the expectation step, we first calculate the exact posterior state probability and joint posterior probability, which can be written as

$$q(\overline{x}_n) = p(\overline{x}_n|\mathbf{S_N}; \mathbf{\Phi}^{i-1}), \tag{13}$$

$$q(\overline{x}_n, \overline{x}_{n-1}) = p(\overline{x}_n, \overline{x}_{n-1}|\mathbf{S_N}; \mathbf{\Phi}^{i-1}), \tag{14}$$

where i is the iteration number. The calculation of (13) and (14) can be performed using forward-backward algorithm [20]. Then, we need to evaluate the posterior Expectation $\mathbb{E}_{\overline{\mathbf{c}}_n|\mathbf{S_N},\overline{x}_n;\mathbf{\Phi}^{i-1}}(\overline{\mathbf{c}}_n)$, which will be used in M-step. By using Bayesian rule and conditional independence property of the proposed HMM model, combining (8), (9) and following the derivation in paper [19], we have

$$q(\overline{\mathbf{c}}_n|\overline{x}_n) = \prod_{f=1}^{F} \mathcal{M}(\overline{c}_{f,n}(1), \cdots, \overline{c}_{f,n}(K); S(f,n), p_{f,n}^{\overline{x}_n}(1), \cdots, p_{f,n}^{\overline{x}_n}(K)), \tag{15}$$

where $\mathcal{M}(\cdot)$ is the multinomial distribution [19].

$$p_{f,n}^{\overline{x}_n}(k) = \frac{\overline{W}_{f,k}^{\overline{x}_n}\overline{H}_{k,n}^{\overline{x}_n}}{\sum_{l=1}^{K}\overline{W}_{f,l}^{\overline{x}_n}\overline{H}_{l,n}^{\overline{x}_n}}. \tag{16}$$

Finally, we have

$$\mathbb{E}(\overline{c}_{f,n}(k)|\mathbf{S_N}, \overline{x}_n) = S(f,n)\frac{\overline{W}_{f,k}^{\overline{x}_n}\overline{H}_{k,n}^{\overline{x}_n}}{\sum_{l=1}^{K}\overline{W}_{f,l}^{\overline{x}_n}\overline{H}_{l,n}^{\overline{x}_n}}. \tag{17}$$

In the maximization step, the purpose is to find parameters to maximize the expected value of complete data likelihood, i.e.,

$$\mathbf{\Phi}^i = \arg\max_{\mathbf{\Phi}} \mathbb{E}_{\overline{\mathbf{X}}_\mathbf{N},\overline{\mathbf{C}}_\mathbf{N}|\mathbf{S_N};\mathbf{\Phi}^{i-1}}[\log p(\mathbf{S_N}, \overline{\mathbf{X}}_\mathbf{N}, \overline{\mathbf{C}}_\mathbf{N})]. \tag{18}$$

By using (18), the estimation of $\overline{\mathbf{A}}$ and $\overline{\boldsymbol{\pi}}$ is the same as the traditional HMM model [20]. To obtain $\overline{\mathbf{W}}^{\overline{x}_n}$ and $\overline{\mathbf{H}}^{\overline{x}_n}$, we can set the derivatives in (18) to zero. Thus, the update of parameters can be written as following:

$$\overline{\pi}_j = \frac{q(\overline{x}_1 = j)}{\sum_{o=1}^{\overline{J}} q(\overline{x}_1 = o)}, \tag{19}$$

$$\overline{A}_{o,j} = \frac{\sum_{n=2}^{\overline{N}} q(\overline{x}_n = j, \overline{x}_{n-1} = o)}{\sum_{j=1}^{\overline{J}}\sum_{n=2}^{\overline{N}} q(\overline{x}_n = j, \overline{x}_{n-1} = o)}, \tag{20}$$

where $1 \le o, j \le \overline{J}$.

$$\overline{\mathbf{W}}^{\overline{x}_n} \leftarrow \overline{\mathbf{W}}^{\overline{x}_n} \odot \frac{\frac{\mathbf{S_N}}{\overline{\mathbf{W}}^{\overline{x}_n}\overline{\mathbf{H}}^{\overline{x}_n}}\mathbf{\Lambda}(j)(\overline{\mathbf{H}}^{\overline{x}_n})^T}{\mathbf{1}(\overline{\mathbf{H}}^{\overline{x}_n})^T}, \tag{21}$$

$$\overline{\mathbf{H}}^{\overline{x}_n} \leftarrow \overline{\mathbf{H}}^{\overline{x}_n} \odot \frac{(\overline{\mathbf{W}}^{\overline{x}_n})^T\frac{\mathbf{S_N}}{\overline{\mathbf{W}}^{\overline{x}_n}\overline{\mathbf{H}}^{\overline{x}_n}}}{(\overline{\mathbf{W}}^{\overline{x}_n})^T\mathbf{1}}, \tag{22}$$

where $\mathbf{\Lambda}(j) = \mathrm{diag}(q(\overline{x}_1 = j), q(\overline{x}_2 = j), \cdots, q(\overline{x}_N = j))$. From (21) and (22), we can find that the parameters update of proposed algorithm is identical to the multiplicative update rule. This ensures that our method is efficient and quick.

3.3. MMSE-based online speech enhancement

In this work, we proposed to combine the NMF-HMM model with MMSE estimator to conduct online speech enhancement. Thus, the estimated signal can be represented as

$$\hat{\mathbf{S}}_n = \mathbb{E}_{\mathbf{S}_n|\mathbf{Y_n}}(\mathbf{S}_n) = \int \mathbf{S}_n p(\mathbf{S}_n|\mathbf{Y_n})\, d\mathbf{S}_n, \tag{23}$$

where $\mathbf{Y_n}$ is defined similar to $\mathbf{S_N}$. We ignore specific details of derivation, the enhanced speech can be represented as

$$\hat{\mathbf{S}}_n = \mathbf{Y}_n \odot \left(\sum_{\overline{x}_n,\ddot{x}_n} \omega_{\overline{x}_n,\ddot{x}_n}\mathbf{p}_n(\overline{x}_n, \ddot{x}_n)\right), \tag{24}$$

where $\omega_{\overline{x}_n,\ddot{x}_n}$ is the weight, which can be written as

$$\omega_{\overline{x}_n,\ddot{x}_n} = \frac{p(\mathbf{Y}_n|\overline{x}_n, \ddot{x}_n)p(\overline{x}_n, \ddot{x}_n|\mathbf{Y_{n-1}})}{\sum_{\overline{x}_n,\ddot{x}_n} p(\mathbf{Y}_n|\overline{x}_n, \ddot{x}_n)p(\overline{x}_n, \ddot{x}_n|\mathbf{Y_{n-1}})}. \tag{25}$$

$$p(\overline{x}_n, \ddot{x}_n|\mathbf{Y_{n-1}}) = \sum_{\overline{x}_{n-1},\ddot{x}_{n-1}} p(\overline{x}_n, \ddot{x}_n|\overline{x}_{n-1}, \ddot{x}_{n-1})p(\overline{x}_{n-1}, \ddot{x}_{n-1}|\mathbf{Y_{n-1}}) \tag{26}$$

In (26), the first term can be acquired by the transition probabilities matrix of noisy speech and the second term is the forward probability that can be calculated by forward algorithm [20]. Additionally, $\mathbf{p}_n(\overline{x}_n, \ddot{x}_n)$ can be represented as

$$\mathbf{p}_n(\overline{x}_n, \ddot{x}_n) = \frac{\overline{\mathbf{W}}^{\overline{x}_n}\overline{\mathbf{H}}^{\overline{x}_n}}{\overline{\mathbf{W}}^{\overline{x}_n}\overline{\mathbf{H}}^{\overline{x}_n} + \ddot{\mathbf{W}}^{\ddot{x}_n}\ddot{\mathbf{H}}^{\ddot{x}_n}}. \tag{27}$$

In enhancement stage, the $\ddot{\mathbf{H}}^{\ddot{x}_n}$ and $\overline{\mathbf{H}}^{\overline{x}_n}$ can be acquired by (5). After that, the enhanced speech can be estimated from (24) to (27). The equation (24) shows that there are more than one basic and activation matrix to be applied to acquire gain to conduct speech enhancement. This is because the proposed algorithm utilize the HMM and consider the temporal aspect. Additionally, the update of activation matrix ($\ddot{\mathbf{H}}^{\ddot{x}_n}$ and $\overline{\mathbf{H}}^{\overline{x}_n}$) can be conducted by parallel computing. This means that our algorithm can reduce the time assumption during the online stage.

4. Experiments and results

4.1. Experimental database preparation

In this study, the proposed algorithm is expected to be evaluated by TIMIT [21] and NOISEX-92 [22] database. During the training stage, all the 4620 utterances from the training TIMIT

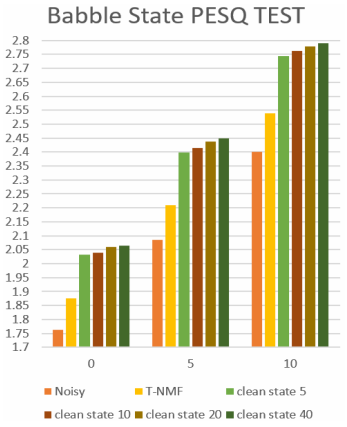

Figure 1: *PESQ score of proposed algorithm in babble noise with different numbers of state.*

database are used to train the clean NMF-HMM model. Additionally, the Babble, F16, Factory and White noise from the NOISEX-92 is also used to train the noise NMF-HMM model. During the test stage, the 200 utterances from the TIMIT test set are randomly chosen to build the test database. Then, four types of noise are added at three different SNR levels (0, 5 and 10dB). The test noise types are F16, Babble, Factory, and White.

In our experiments, all the signals are down-sampled to 16 kHz. The frame length is 1024 samples (64 ms) with a frame shift of 512 samples (32 ms). The size of short time Fourier transform (STFT) is 1024 points with a Hanning window.

4.2. Performance evaluation of speech enhancement

In order to evaluate the performance of the proposed algorithm, there are two test stages. In the first stage, we will investigate the effects of different parameters for NMF-HMM model. This test will be conducted on the babble noise. More specifically, we will investigate the effect of different numbers of state of clean speech for the performance of speech enhancement. In this experiment, the dimension of clean speech and noise mixture is fixed to 25 and 70, respectively, which is based on the previous research [15]. The state of noise is fixed to 2 because we want to show that the proposed algorithm can apply the different noise state to conduct speech enhancement. In this stage, the test result will be evaluated by PESQ [23] and we apply the traditional NMF-based [15] speech enhancement algorithm (T-NMF) as reference method. The aim of this experiment is to acquire the most suitable parameters of NMF-HMM model. Figure 1 shows the experimental result. We can find that the proposed method can achieve the better performance the T-NMF. Additionally, the 40 states for clean speech can achieve the highest score under the all three SNRs. In second stage, the proposed algorithm is expected to be conducted on the more types of noise, which is Babble, F16, Factory and White noise, respectively. We apply the traditional NMF-based [15] speech enhancement algorithm (T-NMF), Optimally-Modified Log-Spectral Amplitude (OM-LSA) method [24] with IMCRA noise estimator [25], linear span filters method [26] (SLF-NMF) that applies the parametric NMF [27] and Log-MMSE [28] algorithm as the reference method. STOI [29] is used to evaluate the performance. For the SLF-NMF, the maximum SNR filter is chosen to conduct the speech enhancement. Furthermore, for the SLF-NMF, the codebook size of clean speech and noise is 64 entries and 8 entries, respectively. The dimension of basic matrix for T-NMF is the same as NMF-HMM. Figure 2 shows the

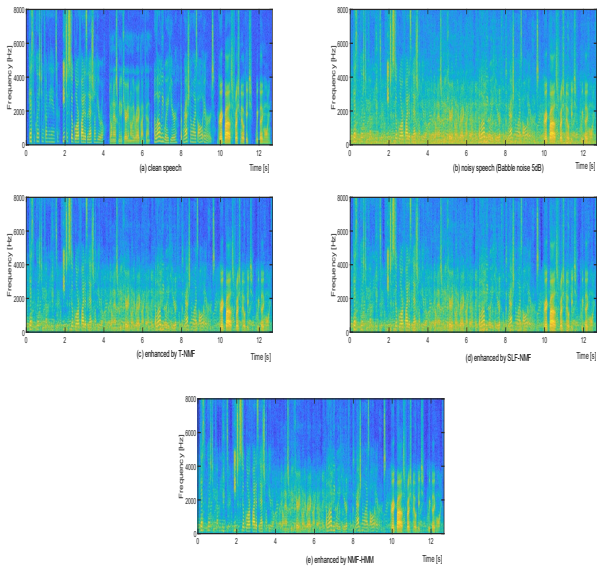

Figure 2: *Spectrum comparison of various NMF-based methods: (a)clean speech, (b)noisy speech with 5dB Babble noise, (c)(d)(e)enhanced speech by T-NMF, SLP-NMF and NMF-HMM,respectivelys*

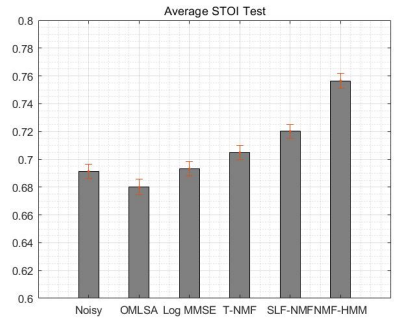

Figure 3: *Average STOI score for four types of noise under three SNRs.*

spectrum comparison of various NMF-based methods. It can be found that the proposed NMF-HMM method is able to remove more noise than other NMF-based method. Meanwhile, NMF-HMM can also recover the more speech information. Figure 3 indicates the average STOI result with the 95% confidential interval (There are four types of noise under three SNRs, each situation includes 200 utterances. Therefore, the average score is acquired by $200\times3\times4$=2400 utterances.) This result shows that NMF-HMM can effectively improve the more speech intelligibility than T-NMF and other reference methods.

5. Conclusions

In this paper, a novel HMM-NMF speech enhancement method is proposed. The core idea is to apply the sum of Poisson as the observation model for each state of HMM because it can ensure that the parameter update rule is identical to the multiplicative update rule. This is quick and efficient. In addition, this method can consider the temporal dynamics of speech signal because of the application of HMM. Furthermore, we proposed a novel HMM-NMF-based MMSE estimator to conduct the online speech enhancement. The experimental results indicate that the proposed algorithm can achieve better speech enhancement performance than these state-of-the-art statistic-based and NMF-based methods.

[1]

6. References

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

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

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[4] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[5] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 700–708, 2003.

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

[7] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

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

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

[10] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[11] 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 Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[12] Y. Xiang and C. Bao, "A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1826–1838, 2020.

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

[14] ——, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[15] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," in *2011 17th International Conference on Digital Signal Processing (DSP)*. IEEE, 2011, pp. 1–6.

[16] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized nonnegative matrix factorization with temporal dependencies for speech denoising," in *Ninth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2008.

[1]This work was supported by Capturi and innovation fund Denmark(No.9065-00046).

[17] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[18] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative Hidden Markov modeling of audio with application to source separation," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 140–148.

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

[20] 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.

[21] 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.

[22] 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 communication*, vol. 12, no. 3, pp. 247–251, 1993.

[23] 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings(ICASSP)*, vol. 2. IEEE, 2001, pp. 749–752.

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

[25] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on speech and audio processing*, vol. 11, no. 5, pp. 466–475, 2003.

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

[27] M. S. Kavalekalam, J. K. Nielsen, L. Shi, M. G. Christensen, and J. Boldt, "Online parametric NMF for speech enhancement," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2320–2324.

[28] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2011.

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