2014 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2014, REIMS, FRANCE

REGULARIZED NMF-BASED SPEECH ENHANCEMENT WITH SPECTRAL COMPONENTS MODELED BY GAUSSIAN MIXTURES

Hanwook Chung[1], *Eric Plourde*[2] *and Benoit Champagne*[1]

[1]Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada
[2]Electrical and Computer Engineering, Sherbrooke University, Sherbrooke, Quebec, Canada
e-mail: {hanwook.chung, benoit.champagne}@mail.mcgill.ca, eric.plourde@usherbrooke.ca

ABSTRACT

In this paper, we introduce a single channel speech enhancement algorithm based on regularized non-negative matrix factorization (NMF). In our proposed formulation, the log-likelihood function (LLF) of the magnitude spectral components, based on Gaussian mixture models (GMM) for both the speech and background noise signals, is included as a regularization term in the NMF cost function. By using this spectral type of regularization, we can incorporate the statistical properties of the signals during the estimation of both the basis and excitation martices in NMF model. Furthermore, borrowing from the expectation-maximization (EM) algorithm and to reduce the computational complexity of the NMF update, the LLF is replaced by its expected value. Experimental results of perceptual evaluation of speech quality (PESQ), source-to-distortion ratio (SDR) and source-to-interference ratio (SIR) show that the proposed speech enhancement algorithm provides better performance than the compared benchmark algorithms.

Index Terms - Regularized non-negative matrix factorization, expectation-maximization, Gaussian mixture model, single channel speech enhancement

1. INTRODUCTION

Speech enhancement algorithms are commonly used to remove additive background noise from a speech signal in order to improve its quality and intelligibility. They have been an attractive research area for decades and now find diverse applications such as in mobile telephony, hearing aid, and speech recognition, to name a few. Despite considerable advances made over the year in this area, the enhancement of speech contaminated by adverse noise, especially under low-SNR or non-stationary conditions, remains an open problem.

Numerous algorithms for single channel speech enhancement have been proposed in the past, such as: spectral subtraction [1], minimum mean-square error (MMSE) estimation [2, 3] and subspace decomposition [4]. Besides these algorithms which use minimal *a priori* information about the speech or noise, further improvements of MMSE-based estimators have been proposed by modeling the speech spectrum as a Rayleigh mixture model (RMM) [5] or a Gaussian mixture model (GMM) [6]. The latter two approaches, which use model parameters derived from a training set for the clean speech, provide a more detailed and accurate description of the speech distribution and are better suited to handle non-stationary speech features. However, the background noise is assumed to be modeled by a single stationary distribution, which is one of the main limitations of these works.

The non-negative matrix factorization (NMF) approach is a method for decomposing a given matrix into basis and excitation matrices with non-negative elements constraint [7]. It has been applied to various problems such as source separation [8, 9], music transcription [10] and speech enhancement [11, 12, 13]. One of the main advantages of NMF-based algorithms in the context of speech enhancement is that they can handle the non-stationarity of both the speech and noise signals simultaneously. In [11], a MMSE filtering method for speech enhancement is proposed where the filter is obtained by using NMF. The temporal dependency of the spectral components between successive time frames is modeled by means of a hidden Markov model (HMM) in [12]. In [13], a methods of speech enhancement based on the use of NMF with priors is proposed.

In order to improve the performance of the NMF, many approaches have been proposed to satisfy certain characteristics of the nature of the signals. A typical approach is to add explicit regularization term to the conventional NMF cost function, and minimize the sum. The use of such additional regularization is considered in [10] and references therein. Although some of the above cited NMF-based techniques exploit regularization through specific mechanisms, such as sparsity and temporal continuity, they do not consider the statistical nature of the signals. In this paper, we introduce a single channel speech enhancement algorithm based on regularized NMF. The novelty of our proposed approach lies in the incorporation of the log-likelihood function (LLF) of the observed data as a regularization term in the NMF cost function. In this formulation, the LLF is derived based on a GMM for both the speech and noise magnitude spectral components. By using this spectral type of regularization, we can exploit the statistical properties of the signals during the estimation of both the basis and excitation matrices. Hence, the proposed algorithm differs from [9] and [13] which apply the statistical model only to the excitation matrix. Furthermore, borrowing from the expectation-maximization (EM) algorithm and to reduce the computational complexity of the NMF update, the LLF is replaced by its expected value. Experimental results of perceptual evaluation of speech quality (PESQ) [19], source-to-distortion ratio (SDR) and source-to-interference ratio (SIR) [20] show that the proposed method provides better performance than the compared benchmark algorithms for speech

[1]Funding for this work was provided by Microsemi Corporation (Ottawa, Canada) and a grant from NSERC (Govt. of Canada).

978-1-4799-3694-6/14/$31.00

enhancement.

2. NMF-BASED SPEECH ENHANCEMENT FRAMEWORK

For a given $K \times L$ matrix $\mathbf{V} = [v_{kl}]$, NMF finds a local optimal decomposition of $\mathbf{V} = \mathbf{WH}$ where $\mathbf{W} = [w_{km}]$ is a $K \times M$ basis matrix, $\mathbf{H} = [h_{ml}]$ is a $M \times L$ excitation matrix and all of these matrices have non-negative elements [7]. The number of basis vectors, M, is typically chosen such that $KM + ML \ll KL$. The factorization is obtained by minimizing a cost function, denoted as $\mathcal{J}(\mathbf{V}, \mathbf{WH})$. By expressing the gradient of the cost function as the difference of two non-negative terms such that $\nabla\mathcal{J}(\mathbf{V}, \mathbf{WH}) = \nabla^{+}\mathcal{J}(\mathbf{V}, \mathbf{WH}) - \nabla^{-}\mathcal{J}(\mathbf{V}, \mathbf{WH})$, solutions can be obtained using general heuristic multiplicative update rules as [9, 10, 14].

$$\begin{cases} \mathbf{W} \leftarrow \mathbf{W} \otimes \dfrac{\nabla^{-}_{\mathbf{W}}\mathcal{J}(\mathbf{V}, \mathbf{WH})}{\nabla^{+}_{\mathbf{W}}\mathcal{J}(\mathbf{V}, \mathbf{WH})} \\ \mathbf{H} \leftarrow \mathbf{H} \otimes \dfrac{\nabla^{-}_{\mathbf{H}}\mathcal{J}(\mathbf{V}, \mathbf{WH})}{\nabla^{+}_{\mathbf{H}}\mathcal{J}(\mathbf{V}, \mathbf{WH})} \end{cases} \tag{1}$$

where the operations $\otimes$ and / respectively denote element-wise multiplication and division. Among various cost functions, the most widely used one is the Kullback-Leibler (KL) divergence, defined as

$$\mathcal{D}_{KL}(\mathbf{V}, \mathbf{WH}) = \sum_{k,l} \left(v_{kl} \ln \frac{v_{kl}}{[\mathbf{WH}]_{kl}} - v_{kl} + [\mathbf{WH}]_{kl} \right) \tag{2}$$

where $[\cdot]_{kl}$ denotes the (k,l)-th entry of its matrix argument. The update rules for the KL divergence are given by [7]

$$\begin{cases} \mathbf{W} \leftarrow \mathbf{W} \otimes \dfrac{(\mathbf{V}/\mathbf{WH})\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T} \\ \mathbf{H} \leftarrow \mathbf{H} \otimes \dfrac{\mathbf{W}^T(\mathbf{V}/\mathbf{WH})}{\mathbf{W}^T\mathbf{1}} \end{cases} \tag{3}$$

where $\mathbf{1}$ is a $K \times L$ matrix of ones and the superscript T a matrix transpose.

In single channel speech enhancement, the observed time-domain signal is decomposed into overlapping frames of length K. The samples in each frame are multiplied by a suitable window function and then transformed to the frequency domain via the short-time Fourier transform (STFT). The resulting signal model is therefore expressed in the time-frequency domain as,

$$Y(k,l) = S(k,l) + N(k,l) \tag{4}$$

where $Y(k,l)$, $S(k,l)$ and $N(k,l)$ respectively denote the STFT of the noisy speech, the clean speech and the additive background noise for the k-th frequency bin of the l-th time frame, where $1 \le k \le K$ and $1 \le l \le L$. Note that the first sample of the STFT coefficients for a given time frame, e.g., $S(1,l)$, denotes the DC component. We assume that the magnitude spectrum of the noisy speech can be approximated by the sum of the clean speech and noise magnitude spectra as $|Y(k,l)| \approx |S(k,l)| + |N(k,l)|$, as it is a practical assumption widely used in the NMF-based speech and audio signal processing [8]-[13]. This signal model can be expressed in a matrix form as $\mathbf{V}_Y \approx \mathbf{V}_S + \mathbf{V}_N = \mathbf{W}_S\mathbf{H}_S + \mathbf{W}_N\mathbf{H}_N$ where the (k,l)-th entry of matrices $\mathbf{V}_Y$, $\mathbf{V}_S$ and $\mathbf{V}_N$ are respectively $v_{Y,kl} = |Y(k,l)|$, $v_{S,kl} = |S(k,l)|$ and $v_{N,kl} = |N(k,l)|$.

NMF-based speech enhancement algorithms consist of two stages. In the training stage, by applying the NMF update rules in (3) to the magnitude spectrum of the training data sets, as represented by matrices $\mathbf{V}_S$ and $\mathbf{V}_N$, the basis matrices, $\mathbf{W}_S$ and $\mathbf{W}_N$, are obtained. In the enhancement stage, these pre-trained basis matrices of clean speech and noise are concatenated as $\mathbf{W}_Y = [\mathbf{W}_S \ \mathbf{W}_N]$. By fixing this basis matrix, the excitation matrix of the noisy speech, $\hat{\mathbf{H}}_Y = [\hat{\mathbf{H}}_S^T \ \hat{\mathbf{H}}_N^T]^T$, is estimated once again by applying the NMF update rule to the magnitude spectrum of the noisy speech, i.e., $\mathbf{V}_Y$. Finally, the magnitude spectrum of the clean speech is estimated $\hat{\mathbf{S}} = \mathbf{W}_S\hat{\mathbf{H}}_S$. The phase of the noisy signal is combined with the estimated magnitude spectrum of the clean speech and is reconstructed in the time domain via the overlap-add method.

3. PROPOSED ALGORITHMS - TRAINING STAGE

In this section, the update rules for the training data sets are introduced. Note that we use general expressions in terms of $\mathbf{W}$ and $\mathbf{H}$ so that the results apply to both the clean speech and noise. The cost function of the regularized NMF is shown as

$$\mathcal{J} = \mathcal{D}(\mathbf{V}, \mathbf{WH}) - \alpha\mathcal{R}(\mathbf{W}, \mathbf{H}) \tag{5}$$

where $\mathcal{D}(\mathbf{V}, \mathbf{WH})$ is a selected measure of the distance between $\mathbf{V}$ and $\mathbf{WH}$, $\alpha > 0$ is a regularization coefficient and $\mathcal{R}(\mathbf{W}, \mathbf{H})$ denotes the regularization term. Note that in this work, a negative sign is applied to the regularization term $\mathcal{R}(\mathbf{W}, \mathbf{H})$ in (5), since the latter will indeed represent a reward (as opposed to a penalty). By expressing the gradient of each term in (5) as the difference of the two positive terms and using the same rules as in (1), the heuristic multiplicative update rules of the regularized NMF can be written as [9, 10],

$$\begin{cases} \mathbf{W} \leftarrow \mathbf{W} \otimes \dfrac{\nabla^{-}_{\mathbf{W}}\mathcal{D}(\mathbf{V}, \mathbf{WH}) + \alpha\nabla^{+}_{\mathbf{W}}\mathcal{R}(\mathbf{W}, \mathbf{H})}{\nabla^{+}_{\mathbf{W}}\mathcal{D}(\mathbf{V}, \mathbf{WH}) + \alpha\nabla^{-}_{\mathbf{W}}\mathcal{R}(\mathbf{W}, \mathbf{H})} \\ \mathbf{H} \leftarrow \mathbf{H} \otimes \dfrac{\nabla^{-}_{\mathbf{H}}\mathcal{D}(\mathbf{V}, \mathbf{WH}) + \alpha\nabla^{+}_{\mathbf{H}}\mathcal{R}(\mathbf{W}, \mathbf{H})}{\nabla^{+}_{\mathbf{H}}\mathcal{D}(\mathbf{V}, \mathbf{WH}) + \alpha\nabla^{-}_{\mathbf{H}}\mathcal{R}(\mathbf{W}, \mathbf{H})} \end{cases} \tag{6}$$

In general, the update rules given in (6) do not guarantee convergence of the iterative process [10]. Especially, a proper value for the regularization coefficient has to be chosen. In our case, this value is determined empirically using validation data (see Section 5).

As for $\mathcal{D}(\mathbf{V}, \mathbf{WH})$, we use the above mentioned KL divergence as given by (2). The gradients of $\mathcal{D}(\mathbf{V}, \mathbf{WH})$ in (6), therefore, are the same as in (3), that is

$$\begin{cases} \nabla^{+}_{\mathbf{W}}\mathcal{D}(\mathbf{V}, \mathbf{WH}) = \nabla^{+}_{\mathbf{W}}\mathcal{D}_{KL}(\mathbf{V}, \mathbf{WH}) = \mathbf{1}\mathbf{H}^T \\ \nabla^{-}_{\mathbf{W}}\mathcal{D}(\mathbf{V}, \mathbf{WH}) = \nabla^{-}_{\mathbf{W}}\mathcal{D}_{KL}(\mathbf{V}, \mathbf{WH}) = (\mathbf{V}/\mathbf{WH})\mathbf{H}^T \\ \nabla^{+}_{\mathbf{H}}\mathcal{D}(\mathbf{V}, \mathbf{WH}) = \nabla^{+}_{\mathbf{H}}\mathcal{D}_{KL}(\mathbf{V}, \mathbf{WH}) = \mathbf{W}^T\mathbf{1} \\ \nabla^{-}_{\mathbf{H}}\mathcal{D}(\mathbf{V}, \mathbf{WH}) = \nabla^{-}_{\mathbf{H}}\mathcal{D}_{KL}(\mathbf{V}, \mathbf{WH}) = \mathbf{W}^T(\mathbf{V}/\mathbf{WH}) \end{cases} \tag{7}$$

The magnitude spectrum of the signal can be modeled by a GMM to deal with non-stationarities [6]. Under this statistical assumption, we use the corresponding LLF as the regularization term

in (5). Note that this approach differs from [9, 13] which apply the statistical model only to the excitation matrix.

Since $\mathbf{V} \approx \mathbf{WH}$, we consider $\mathbf{WH}$ as the observation matrix where the probability density function (pdf) of each column, say $[\mathbf{WH}]_l$ for $l \in \{1, ...L\}$, is modeled by a Gaussian mixture. By using this assumption, we can derive the gradients of the regularization term with respect to $\mathbf{W}$ and $\mathbf{H}$. For a K-dimensional multi-variate random vector $[\mathbf{WH}]_l$, the GMM is defined in terms of the parametric probabilistic model

$$f([\mathbf{WH}]_l|\boldsymbol{\theta}) = \sum_{i=1}^{I} m_i \mathcal{N}([\mathbf{WH}]_l|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{8}$$

where I is the number of Gaussian mixture components, m_i for $i \in \{1, \ldots, I\}$ are mixing coefficients such that $\sum_i m_i = 1$, $\mathcal{N}([\mathbf{WH}]_l|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes the pdf of K-dimensional Gaussian pdf with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ for the i-th component and $\boldsymbol{\theta} = \{m_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{I}$ is a parameter set of a GMM. Under the assumption of independence for STFT data obtained over different frames, the LLF of the observed set $\mathbf{WH}$ is given by

$$\mathcal{L}(\mathbf{WH}|\boldsymbol{\theta}) = \sum_{l=1}^{L} \ln f([\mathbf{WH}]_l|\boldsymbol{\theta}). \tag{9}$$

Recall that matrix $\mathbf{WH}$ represents the magnitude spectrum of the training signal, either the clean speech or noise in our case. By using Jensen's inequality, we can construct a lower bound on the LLF such that [15],

$$\begin{aligned}\mathcal{L}(\mathbf{WH}|\boldsymbol{\theta}) &\geq \sum_{l=1}^{L}\sum_{i=1}^{I} P_r(z_i{=}1|[\mathbf{WH}]_l) \ln \frac{m_i \mathcal{N}([\mathbf{WH}]_l|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{P_r(z_i{=}1|[\mathbf{WH}]_l)} \\ &\triangleq \quad \mathcal{L}_B(\mathbf{WH}|\boldsymbol{\theta})\end{aligned} \tag{10}$$

where $\mathbf{z} = [z_1, z_2, \ldots, z_I]^T$ is a I-dimensional binary latent vector in which $z_i \in \{0, 1\}$ and $\sum_i z_i = 1$. Each element, z_i, is an indicator of the corresponding Gaussian component whose marginal distribution is related to the corresponding mixing coefficient as in $m_i = P_r(z_i = 1)$ and $\mathcal{L}_B(\mathbf{WH}|\boldsymbol{\theta})$ denotes the lower bound of the LLF. $P_r(z_i = 1|[\mathbf{WH}]_l)$ is the posterior probability of $z_i = 1$ given the observation $[\mathbf{WH}]_l$, which is given by,

$$P_r(z_i = 1|[\mathbf{WH}]_l) = \frac{m_i \mathcal{N}([\mathbf{WH}]_l|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{i=1}^{I} m_i \mathcal{N}([\mathbf{WH}]_l|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}. \tag{11}$$

In the expectation-maximization (EM) algorithm, the posterior probability is computed during the expectation step (E step), and the parameters are estimated by maximizing $\mathcal{L}_B(\mathbf{WH}|\boldsymbol{\theta})$ in the maximization step (M step). Since the posterior probability is fixed during the M step, the maximization problem is equivalent to maximizing the expected value of the complete-data LLF with respect to the posterior distribution of the latent vector, $\mathcal{L}_C(\mathbf{WH}|\boldsymbol{\theta})$, which is defined as

$$\begin{aligned}\mathcal{L}_C(\mathbf{WH}|\boldsymbol{\theta}) \triangleq \sum_{l=1}^{L}\sum_{i=1}^{I} & P_r(z_i = 1|[\mathbf{WH}]_l) \\ & \cdot \ln[m_i \mathcal{N}([\mathbf{WH}]_l|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)]\end{aligned} \tag{12}$$

Instead of using the LLF described in (9) as in [9], we propose to use $\mathcal{L}_C(\mathbf{WH}|\boldsymbol{\theta})$ as the regularization term in (5) to lower the computational complexity, that is $\mathcal{R}(\mathbf{W}, \mathbf{H}) = \mathcal{L}_C(\mathbf{WH}|\boldsymbol{\theta})$.

We considered diagonal-variance by ignoring the correlations between spectral components for simplicity. The components of the gradients of the regularization terms are found as

$$\nabla_{w_{km}} \mathcal{L}_C = \nabla^{+}_{w_{km}} \mathcal{L}_C - \nabla^{-}_{w_{km}} \mathcal{L}_C \tag{13}$$

where

$$\nabla^{+}_{w_{km}} \mathcal{L}_C = \sum_{l=1}^{L}\sum_{i=1}^{I} P_r(z_i = 1|[\mathbf{WH}]_l) \Sigma^{-1}_{i,kk} \mu_{i,k} h_{ml} \tag{14}$$

$$\nabla^{-}_{w_{km}} \mathcal{L}_C = \sum_{l=1}^{L}\sum_{i=1}^{I} P_r(z_i = 1|[\mathbf{WH}]_l) \Sigma^{-1}_{i,kk} [\mathbf{WH}]_{kl} h_{ml} \tag{15}$$

and

$$\nabla_{h_{ml}} \mathcal{L}_C = \nabla^{+}_{h_{ml}} \mathcal{L}_C - \nabla^{-}_{h_{ml}} \mathcal{L}_C \tag{16}$$

where

$$\nabla^{+}_{h_{ml}} \mathcal{L}_C = \sum_{k=1}^{K}\sum_{i=1}^{I} P_r(z_i = 1|[\mathbf{WH}]_l) \Sigma^{-1}_{i,kk} \mu_{i,k} w_{km} \tag{17}$$

$$\nabla^{-}_{h_{ml}} \mathcal{L}_C = \sum_{k=1}^{K}\sum_{i=1}^{I} P_r(z_i = 1|[\mathbf{WH}]_l) \Sigma^{-1}_{i,kk} [\mathbf{WH}]_{kl} w_{km} \tag{18}$$

where $\mu_{i,k}$ denotes the k-th element of the mean vector $\boldsymbol{\mu}_i$ and $\Sigma^{-1}_{i,kk}$ indicates the (k, k)-th element of the inverse covariance matrix $\boldsymbol{\Sigma}_i^{-1}$. The dependence of $\mathcal{L}_C(\mathbf{WH}|\boldsymbol{\theta})$ in (13)-(18) on $\mathbf{WH}$ and $\boldsymbol{\theta}$ is omitted for notational convenience. Since the posterior probability and all elements of the mean vector and covariance matrix are non-negative, the values from (14), (15), (17) and (18) will be non-negative. By using (6), therefore, $\mathbf{W}$ and $\mathbf{H}$ are updated under the non-negative elements constraint.

The parameter set $\boldsymbol{\theta} = \{m_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{I}$ can be estimated by employing the EM algorithm as in [15], which results in

$$\begin{aligned} m_i &= \frac{1}{L}\sum_{l=1}^{L} P_r(z_i = 1|[\mathbf{WH}]_l) \\ \mu_{i,k} &= \frac{\sum_{l=1}^{L} P_r(z_i = 1|[\mathbf{WH}]_l)[\mathbf{WH}]_{kl}}{\sum_{l=1}^{L} P_r(z_i = 1|[\mathbf{WH}]_l)} \\ \Sigma_{i,kk} &= \frac{\sum_{l=1}^{L} P_r(z_i = 1|[\mathbf{WH}]_l)([\mathbf{WH}]_{kl} - \mu_{i,k})^2}{\sum_{l=1}^{L} P_r(z_i = 1|[\mathbf{WH}]_l)} \end{aligned} \tag{19}$$

In our approach, the initialization of $\mathbf{W}$ and $\mathbf{H}$ is performed by running the conventional NMF with KL-divergence constraint while the magnitude spectrum of the signal, $\mathbf{V}$, is used to train the initial parameter set, $\boldsymbol{\theta}$, via the EM algorithm. Each algorithm is also initialized by using positive random numbers and k-means clustering.

Two different training algorithms are considered. We first propose to use a two-stage joint update (TSJU) method which follows a similar approach as the joint update method described in [9]. At

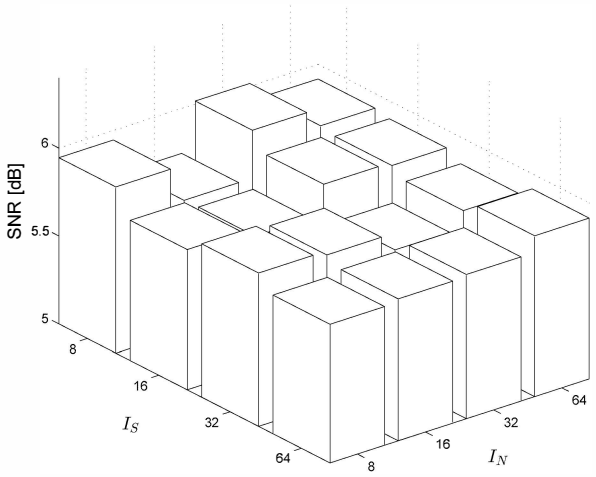

Fig. 1. Average SNR values of estimated clean speech with different numbers of components in the GMM.

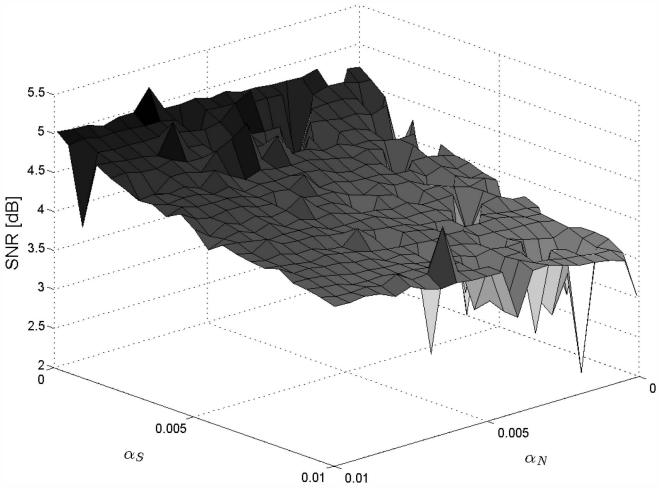

Fig. 2. Average SNR values of estimated clean speech from babble noise with different regularization coefficients for the enhancement stage.

each iteration: the posterior distribution is first calculated using (11); $\mathbf{W}$ and $\mathbf{H}$ are then updated using (6), (7), (14), (15), (17) and (18); at the end of the iteration, $\boldsymbol{\theta}$ is estimated by running the EM algorithm.

Focusing on the regularization term $\mathcal{R}(\mathbf{W}, \mathbf{H})$ in (5) where we used the expected LLF, we can consider the update rules for $\mathbf{W}$ and $\mathbf{H}$ along with $\boldsymbol{\theta}$ as a parameter estimation using the EM algorithm. Based on this idea, we suggest another algorithm which will be referred to as the single-stage joint update (SSJU) method and where all the parameters, $\mathbf{W}$, $\mathbf{H}$ and $\boldsymbol{\theta}$, are estimated in one stage. At each iteration: the posterior distribution is first calculated using (11); $\mathbf{W}$ and $\mathbf{H}$ are then updated using (6), (7), (14), (15), (17) and (18); at the end of the iteration, $\boldsymbol{\theta}$ is calculated using (19). Note that SSJU differs from TSJU, which uses an additional iterative computation for the EM algorithm to estimate $\boldsymbol{\theta}$ at the end of each iteration. Therefore, a more efficient implementation is provided by SSJU.

4. PROPOSED ALGORITHMS - ENHANCEMENT STAGE

The enhancement stage follows a similar strategy as in [9]. First, the basis matrices of both clean speech and noise are concatenated as $\mathbf{W}_Y = [\mathbf{W}_S \ \mathbf{W}_N]$. Once the magnitude spectrum of the noisy speech is computed, the excitation matrix of the noisy speech, $\hat{\mathbf{H}}_Y = [\hat{\mathbf{H}}_S^T \ \hat{\mathbf{H}}_N^T]^T$, is estimated by fixing the basis matrix $\mathbf{W}_Y$ and the parameter sets of both clean speech and noise, $\boldsymbol{\theta}_S = \{m_{i,S}, \boldsymbol{\mu}_{i,S}, \boldsymbol{\Sigma}_{i,S}\}_{i=1}^{I_S}$ and $\boldsymbol{\theta}_N = \{m_{i,N}, \boldsymbol{\mu}_{i,N}, \boldsymbol{\Sigma}_{i,N}\}_{i=1}^{I_N}$, which are estimated during the training stage. The cost function in the enhancement stage can be written as

$$\mathcal{J} = D_{KL}(\mathbf{V}_Y, \mathbf{W}_Y\mathbf{H}_Y) - \mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y) \tag{20}$$

$$\mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y) = \alpha_S \mathcal{L}_C(\mathbf{W}_S\mathbf{H}_S|\boldsymbol{\theta}_S) + \alpha_N \mathcal{L}_C(\mathbf{W}_N\mathbf{H}_N|\boldsymbol{\theta}_N) \tag{21}$$

where $\mathcal{L}_C(\mathbf{W}_S\mathbf{H}_S|\boldsymbol{\theta}_S)$ and $\mathcal{L}_C(\mathbf{W}_S\mathbf{H}_S|\boldsymbol{\theta}_S)$ take the same form as in (12), α_S and α_N are regularization coefficients of the clean speech and noise for the enhancement stage, respectively. The excitation matrix is obtained in a similar way as in (6), that is

$$\mathbf{H}_Y \leftarrow \mathbf{H}_Y \otimes \frac{\nabla_{\mathbf{H}}^- \mathcal{D}_{KL}(\mathbf{V}_Y, \mathbf{W}_Y\mathbf{H}_Y) + \nabla_{\mathbf{H}}^+ \mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y)}{\nabla_{\mathbf{H}}^+ \mathcal{D}_{KL}(\mathbf{V}_Y, \mathbf{W}_Y\mathbf{H}_Y) + \nabla_{\mathbf{H}}^- \mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y)} \tag{22}$$

The gradient terms in (22) are given by (7), (17) and (18).

As in the conventional NMF, the magnitude spectrum of the clean speech is estimated by $\hat{\mathbf{S}} = \mathbf{W}_S\hat{\mathbf{H}}_S$. Finally, the phase of the noisy speech is combined with the estimated magnitude spectrum of the clean speech and the enhanced speech signal is reconstructed in time domain.

5. EXPERIMENTS

5.1. Methodology

In this section, experiments and performance evaluations of our methods are presented. We used clean speech from the TSP database [16] and noise from the NOISEX database [17], where the sampling rate of all signals was adjusted to 8 kHz. The noisy signals were generated by adding babble noise and factory noise at three different input signal-to-noise ratios (SNR) of 0, 5 and 10 dB. Magnitude spectrum of each signal is obtained by using Hanning window of 512 samples with 50 % overlap. The signal synthesis was performed using the overlap-add method.

For clean speech, 8 speakers (4 male and 4 female) were considered with 10 sentences each for training data, 2 sentences for validation and test data. For babble and factory noise, 30 second-long signals were used for training set and a length which corresponds to the validation and test data of the clean speech was used for validation and test sets. Note that all these data sets were disjoint. Validation and test sets were used to generate noisy speech. We examined the algorithms with 96 basis vectors for both clean speech and noise. Besides the number of the basis vectors, appropriate values of hyper-parameters such as the number of components in a GMM and regularization coefficients had to be obtained.

The hyper-parameters in our model are the number of components in the GMM, I, the regularization coefficients for the training stage, α_{train}, and the enhancement stage, α_S and α_N. The vali-

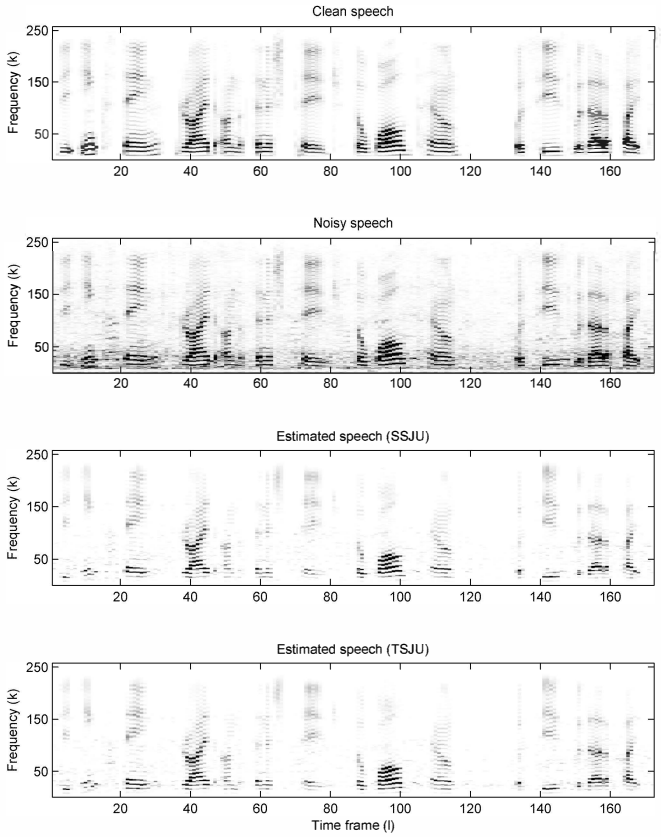

Fig. 3. Example of magnitude spectrum of clean, noisy and estimated clean speech. A male speech is degraded with babble noise at 5 dB input SNR.

dation sets were used for the estimation of hyper-parameters. First, the regularization coefficient in the training stage was obtained by observing the convergence behavior of the KL divergence in (2) as a measure of the decomposition error. We found that the proposed training method converged well for values of α_{train} under 0.0003. In our implementation, we used $\alpha_{train} = 0.00005$ for both the clean speech and noise. After fixing α_{train}, we conducted simulations for different number of mixtures in the GMM, i.e., $I \in \{8, 16, 32, 64\}$. Fig. 1 shows the average SNR of the estimated clean speech using the SSJU approach by varying $\alpha_S = \alpha_N$ from 0.0001 to 0.001 with 0.0001 increments. The highest SNR was found at $I_S = 8$ and $I_N = 32$ and consequently, we used these values in our implementation of the GMM. The same were found when using the TSJU approach. Finally, appropriate values for α_S and α_N were obtained by fixing the previously estimated hyper-parameters α_{train}, I_S and I_N. An example of average SNR values of estimated clean speech from babble noise with different α_S and α_N is shown in Fig. 2. A similar pattern was also found with factory noise for both SSJU and TSJU. Based on this observation, we chose $(\alpha_S, \alpha_S) = (0.0001, 0.008)$.

5.2. Results

We used perceptual evaluation of speech quality (PESQ) [19], source-to-distortion ratio (SDR) and source-to-interference ratio (SIR) [20] as the objective measures of performance. The PESQ attempts to predict the overall perceptual quality in mean opinion score (MOS) yielding a result from 1 to 4.5. The SDR reflects the overall separation quality on a dB scale, considering both speech distortion and noise reduction aspects. The SIR reflects the noise reduction also measured on a dB scale. For all these measures, a higher value indicates a better result. As for the comparison, we considered Wiener filtering method, conventional NMF (CNMF) and the method introduced by Grais et al.[9], which will be referred to as RNMF. The Wiener filter was obtained by using the noise power spectral density (PSD) estimation introduced in [18]. Hyper-parameters in RNMF were obtained using a similar process as the one described in Sec. 5.1.

Table 1. PESQ, SDR and SIR for babble noise

Input SNR	Eval.	Wiener	CNMF	RNMF	SSJU	TSJU
0 dB	PESQ	1.74	1.72	1.77	**1.9**	1.89
	SDR	0.82	1.61	2.3	**3.98**	3.88
	SIR	1.98	3.96	1.95	**7.5**	7.2
5 dB	PESQ	2.07	2.09	2.08	**2.24**	**2.24**
	SDR	5.76	5.88	6.76	7.9	**8.01**
	SIR	7.31	9.8	7.07	**15.66**	15.34
10 dB	PESQ	2.31	2.44	2.41	**2.48**	2.47
	SDR	9.82	8.85	**9.87**	9.35	9.56
	SIR	12.32	15.06	12.21	**22.16**	21.8

Table 2. PESQ, SDR and SIR for factory noise

Input SNR	Eval.	Wiener	CNMF	RNMF	SSJU	TSJU
0 dB	PESQ	1.76	1.8	1.76	**1.97**	**1.97**
	SDR	3.96	4.36	3.99	5.9	**5.96**
	SIR	7.68	6.58	4.55	9.93	**10.2**
5 dB	PESQ	2.08	2.14	2.09	2.3	**2.32**
	SDR	8.01	8.73	8.62	9.74	**9.78**
	SIR	12.94	11.85	9.74	17.42	**17.67**
10 dB	PESQ	2.33	2.46	2.38	2.53	**2.54**
	SDR	10.98	12.13	**12.14**	11.49	11.58
	SIR	17.65	16.87	14.76	23.34	**23.6**

Fig. 3 demonstrates an example of the magnitude spectrums of the estimated clean speech using the proposed SSJU and TSJU methods. In this particular example, a male speech is degraded with babble noise at 5 dB input SNR. The bottom spectrographs clearly show that the proposed methods have been able to significantly reduce the background noise. For this case, we obtained (PESQ, SDR, SIR) = (2.45, 8.40, 14.65) and (2.52, 8.53, 14.36) for SSJU and TSJU, respectively. Table 1 and 2 show the objective results of the proposed algorithms and reference methods for babble and factory noises at different SNR levels. It can be seen that in most cases (except for SDR at 10dB), the proposed SSJU and TSJU algorithms perform better than the reference methods. In particular, both algorithms lead to significantly higher values of PESQ and SIR. Moreover, the proposed algorithms showed significantly reduced noise in the estimated clean speech based on the SIR observation, while maintaining the best PESQ scores. Another interesting observation is that the proposed methods SSJU and TSJU provided similar results. This indicates that highly improved performance can be obtained by using SSJU which is more efficient than TSJU in terms of computational

complexity.

Informal listening experiments were also conducted to evaluate the performance of the proposed algorithms. It was generally found that SSJU and TSJU offered the best performance, followed by CNMF, RNMF and Wiener in that order. In terms of noise reduction, the proposed algorithms could remove more of the additive background acoustic noise than the other NMF-based algorithms under evaluation, which in turn all performed better than Wiener filtering, where the residual noise exhibited a less natural character. In terms of speech distortion, it was found that the enhanced speech was slightly more clearly audible with the proposed SSJU and TSJU algorithms, although some of the low frequencies in the voice sounds were somewhat attenuated. We conjecture that this might be corrected by the use of spectral weighting in the NMF cost function or the associated regularization; this remains an interesting avenue for future work.

6. CONCLUSIONS

A single channel speech enhancement algorithm based on regularized NMF has been proposed. Magnitude spectral components of both clean speech and noise were modeled by GMM. The corresponding LLF was used as the regularization to the cost function of the conventional NMF in order to exploit the statistical characteristics of the signals. The basis and excitation matrices were estimated using multiplicative update rules under the proposed regularization. In addition, the expected value of the LLF was employed for the regularization so that more efficient update equations are obtained. Experimental results based on PESQ, SDR and SIR showed that the proposed algorithm improves the performance.

7. REFERENCES

[1] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system" *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 126-137, Mar. 1999.

[2] 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, Dec. 1984.

[3] E. Plourde and B. Champagne, "Auditory-based spectral amplitude estimators for speech enhancement," *IEEE Trans. Audio, Speech and Language Process.*, vol. 16, no. 8, pp. 1614-1623, Nov. 2008.

[4] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 4, pp. 334-341, July 2003.

[5] J. S. Erkelens, J. Jensen and R. Heusdens, "Speech enhancement based on Rayleigh mixture modeling of speech spectral amplitude distributions," in *Proc. EUSIPCO*, pp. 65-69, Sept. 2007.

[6] A. Kundu, S. Chatterjee, A. S. Murthy and T. V. Sreenivas, "GMM based Bayesian approach to speech enhancement in signal/transform domain," in *Proc. ICASSP*, pp. 4893-4896, Apr., 2008.

[7] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, pp. 556-562, 2001.

[8] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech and Language Process.*, vol. 15, pp. 1-12, Jan. 2007.

[9] E. M. Grais and H. Erdogan, "Regularized nonnegative matrix factorization using Gaussian mixture priors for unsupervised single channel source separation," *Computer Speech and Language*, vol. 27, no. 3, pp. 746-762, May 2013.

[10] N. Bertin, R. Badeu and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech and Language Process.*, vol. 18, no. 3, pp, 538-549, Mar. 2010.

[11] N. Mohammadiha, T. Gerkmann and A. Leijon, "A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization," in *Proc. WASPAA*, pp. 45-48, Oct. 2011.

[12] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. ICASSP*, pp. 17-20, May 2011.

[13] K. W. Wilson, B. Raj, P. Smaragdis and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. ICASSP*, pp. 4029-4032, Apr. 2008.

[14] C. Fevotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, pp. 2421-2456, Sept. 2011.

[15] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[16] P. Kabal, "TSP speech database," Tech. Rep., McGill University, Montreal, Canada, 2002.

[17] Rice University, "Signal processing information base: noise data," Available online: http://spib.rice.edu/spib/select_noise.html.

[18] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio Speech and Language Process.*, vol. 20, no. 4, pp. 1383-1393, May 2012.

[19] ITU-T. P.862, "Perceptual evaluation of speech quality (PESQ): and objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Tech. Rep., 2000.

[20] E. Vincent, R. Gribonval and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech and Language Process.*, vol. 14, no. 4, pp. 1462-1469, July 2006.