Spectro-temporal post-enhancement using MMSE estimation in NMF based single-channel source separation

Emad M. Grais and Hakan Erdogan

Faculty of Engineering and Natural Sciences
Sabanci University, Orhanli Tuzla, 34956, Istanbul, Turkey.
grais@sabanciuniv.edu, haerdogan@sabanciuniv.edu

Abstract

We propose to use minimum mean squared error (MMSE) estimates to enhance the signals that are separated by nonnegative matrix factorization (NMF). In single channel source separation (SCSS), NMF is used to train a set of basis vectors for each source from their training spectrograms. Then NMF is used to decompose the mixed signal spectrogram as a weighted linear combination of the trained basis vectors from which estimates of each corresponding source can be obtained. In this work, we deal with the spectrogram of each separated signal as a 2D distorted signal that needs to be restored. A multiplicative distortion model is assumed where the logarithm of the true signal distribution is modeled with a Gaussian mixture model (GMM) and the distortion is modeled as having a log-normal distribution. The parameters of the GMM are learned from training data whereas the distortion parameters are learned online from each separated signal. The initial source estimates are improved and replaced with their MMSE estimates under this new probabilistic framework. The experimental results show that using the proposed MMSE estimation technique as a post enhancement after NMF improves the quality of the separated signal.

Index Terms: Single channel source separation, nonnegative matrix factorization, Minimum mean square error estimates, and Gaussian mixture models.

1. Introduction

In single channel source separation problems, only one observation of the mixed signal is available. The solution of this problem usually relies on training data for each source signal. Nonnegative matrix factorization (NMF) [1] is usually used to train a set of basis vectors (basis matrix) for each source signal. NMF is then used to decompose the mixed signal spectrogram as a weighted linear combination of the trained basis matrices for all sources in the mixed signal. The estimate for each source is computed by summing the decomposition terms that include its corresponding trained basis vectors [2, 3]. The trained basis matrix is used as the only representative for the training data for each source. The trained basis matrices are then used in mixed signal decomposition in the separation/testing stage.

The trained basis matrix that is usually used as the only representative for each source training data is usually not sufficient to represent all the characteristics of each source. This representation may be limited since the dynamic information between frames is missing and there is no analytical approach for choosing a suitable number of bases. More information about the sources besides their trained basis matrices is usually needed.

This work was supported by Turk-Telekom under grant number 3014-06.

In this work, we support the NMF based source separation with source enhancement for the separated signal. Besides training a basis matrix for each source, the spectrogram for each training data is directly used to train a Gaussian mixture model (GMM) in the logarithm domain. The trained basis matrices are used with NMF to find a separated signal for each source in the mixed signal. The spectrogram of each separated signal is then treated as a 2D distorted signal. The trained GMMs and the expectation maximization algorithm (EM) [4] are used to learn the distortion in each separated signal spectrogram. The trained GMMs, the learned distortion, the minimum mean square error (MMSE) estimates, and the Wiener filters are used to find enhanced versions of the separated signals. To consider the dynamic information between the spectrogram frames, we apply the enhancement approach on multiple consequent frames at once instead of applying it frame by frame.

This paper is organized as follows: In Section 2, a brief introduction about NMF is presented. Section 3 describes SCSS problem and the conventional approach for using NMF in SCSS. In Section 4, we introduce the MMSE estimation based post enhancement for the separated source signals which is our main contribution in this paper. In the remaining sections we present our experimental results.

2. Nonnegative matrix factorization

Nonnegative matrix factorization is a matrix factorization algorithm that decomposes any nonnegative matrix $\boldsymbol{V}$ into a multiplication of a nonnegative basis matrix $\boldsymbol{B}$ and a nonnegative gains matrix $\boldsymbol{G}$ as follows:

$$\boldsymbol{V} \approx \boldsymbol{B}\boldsymbol{G}. \tag{1}$$

The matrix $\boldsymbol{B}$ contains the basis vectors that are optimized to allow the data in $\boldsymbol{V}$ to be approximated as a linear combination of its constituent columns. The solution for $\boldsymbol{B}$ and $\boldsymbol{G}$ can be found by minimizing the following Itakura-Saito (IS) divergence cost function [5]:

$$\min_{\boldsymbol{B},\boldsymbol{G}} D_{IS}\left(\boldsymbol{V} \,||\, \boldsymbol{B}\boldsymbol{G}\right), \tag{2}$$

where

$$D_{IS}\left(\boldsymbol{V} \,||\, \boldsymbol{B}\boldsymbol{G}\right) = \sum_{a,b} \left(\frac{\boldsymbol{V}_{a,b}}{(\boldsymbol{B}\boldsymbol{G})_{a,b}} - \log \frac{\boldsymbol{V}_{a,b}}{(\boldsymbol{B}\boldsymbol{G})_{a,b}} - 1 \right).$$

This divergence cost function is a good measurement for the perceptual difference between different audio signals [5]. The IS-NMF solution for equation (2) can be computed by alternating multiplicative updates of $\boldsymbol{G}$ and $\boldsymbol{B}$ as follows:

$$\boldsymbol{G} \leftarrow \boldsymbol{G} \otimes \frac{\boldsymbol{B}^T \left(\frac{\boldsymbol{V}}{(\boldsymbol{B}\boldsymbol{G})^2} \right)}{\boldsymbol{B}^T \left(\frac{\boldsymbol{1}}{\boldsymbol{B}\boldsymbol{G}} \right)}, \tag{3}$$

$$\boldsymbol{B} \leftarrow \boldsymbol{B} \otimes \frac{\left(\frac{\boldsymbol{V}}{(\boldsymbol{BG})^2}\right)\boldsymbol{G}^T}{\left(\frac{\boldsymbol{1}}{\boldsymbol{BG}}\right)\boldsymbol{G}^T}, \tag{4}$$

where $\boldsymbol{1}$ is a matrix of ones with the same size of $\boldsymbol{V}$, the operation $\otimes$ is an element-wise multiplication, all divisions and $(.)^2$ are element-wise operations. The matrices $\boldsymbol{B}$ and $\boldsymbol{G}$ are usually initialized by positive random numbers and then updated iteratively using equations (3) and (4).

3. Single channel source separation

In SCSS problems, the aim is to find estimates of source signals that are mixed on a single recording $y(t)$. In this work, we assume the number of sources is two. This problem is usually solved in the short time Fourier transform (STFT) domain. Let $Y(t,f)$ be the STFT of $y(t)$, where t represents the frame index and f is the frequency-index. Due to the linearity of the STFT, we have

$$Y(t,f) = S_1(t,f) + S_2(t,f), \tag{5}$$

where $S_1(t,f)$ and $S_2(t,f)$ are the unknown STFT of the sources in the mixed signal. Assuming independence of the sources, we can write the power spectral density (PSD) of the measured signal as the sum of source signal PSDs as follows:

$$\sigma_y^2(t,f) = \sigma_1^2(t,f) + \sigma_2^2(t,f). \tag{6}$$

We can write the PSDs in matrix form (power spectrograms) as follows:

$$\boldsymbol{Y} = \boldsymbol{S}_1 + \boldsymbol{S}_2, \tag{7}$$

where $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are the unknown spectrograms of the source signals, and they need to be estimated using the observed mixed signal spectrogram $\boldsymbol{Y}$ and the training data for each source. The PSD for the measured signal $y(t)$ is calculated by taking the squared magnitude of the DFT of the windowed signal.

The main idea to solve for $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ is to use NMF to train a set of basis vectors for each source signal. NMF trains the source bases for each source by decomposing the power spectrogram of its corresponding training data as follows:

$$\boldsymbol{S}_1^{train} \approx \boldsymbol{B}_1\boldsymbol{G}_1^{train}, \quad \boldsymbol{S}_2^{train} \approx \boldsymbol{B}_2\boldsymbol{G}_2^{train}, \tag{8}$$

where $\boldsymbol{S}_1^{train}$ and $\boldsymbol{S}_2^{train}$ are the spectrograms of the training data for the first and second source respectively, the columns of $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are considered as trained bases that are used in mixed signal decomposition as shown in next sections. The update rules in equations (4) and (3) are used to decompose $\boldsymbol{S}_1^{train}$ and $\boldsymbol{S}_2^{train}$ in equation (8). After each NMF iteration the columns in the basis matrices are normalized using the ℓ^2 norm and the gain matrices are calculated accordingly.

After observing the mixed signal, NMF is used to decompose the mixed signal spectrogram $\boldsymbol{Y}$ with the trained basis matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ for the first and second source respectively as follows:

$$\boldsymbol{Y} \approx [\boldsymbol{B}_1, \boldsymbol{B}_2]\,\boldsymbol{G} \quad \text{or} \quad \boldsymbol{Y} \approx [\boldsymbol{B}_1 \quad \boldsymbol{B}_2]\begin{bmatrix} \boldsymbol{G}_1 \\ \boldsymbol{G}_2 \end{bmatrix}. \tag{9}$$

The only unknown here is the gains matrix $\boldsymbol{G}$ which can be calculated iteratively using the update rule in equation (3). The basis matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ were trained as shown in equation (8) and they are fixed in this separation stage. The initial spectrogram estimate for each source can be computed as follows:

$$\widetilde{\boldsymbol{S}}_1 = \boldsymbol{B}_1\boldsymbol{G}_1, \qquad \widetilde{\boldsymbol{S}}_2 = \boldsymbol{B}_2\boldsymbol{G}_2. \tag{10}$$

The initial estimated spectrograms $\widetilde{\boldsymbol{S}}_1$ and $\widetilde{\boldsymbol{S}}_2$ are used to build spectral masks (Wiener filter) [5, 6] as follows:

$$\boldsymbol{H}_1 = \frac{\widetilde{\boldsymbol{S}}_1}{\widetilde{\boldsymbol{S}}_1 + \widetilde{\boldsymbol{S}}_2}, \qquad \boldsymbol{H}_2 = \frac{\widetilde{\boldsymbol{S}}_2}{\widetilde{\boldsymbol{S}}_1 + \widetilde{\boldsymbol{S}}_2}, \tag{11}$$

where the divisions are done element-wise. The final estimate of each source STFT can be obtained as follows:

$$\hat{S}_1(t,f) = \boldsymbol{H}_1(t,f)\,Y(t,f), \tag{12}$$

$$\hat{S}_2(t,f) = \boldsymbol{H}_2(t,f)\,Y(t,f), \tag{13}$$

where $Y(t,f)$ is the STFT of the observed mixed signal in equation (5), $\boldsymbol{H}_1(t,f)$ and $\boldsymbol{H}_2(t,f)$ are the entries at row f and column t of the spectral masks $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ respectively. The spectral mask entries scale the observed mixed signal STFT entries according to the contribution of each source in the mixed signal [3, 7, 8]. The estimated source signals $\hat{s}_1(t)$ and $\hat{s}_2(t)$ can be found by inverse STFT of $\hat{S}_1(t,f)$ and $\hat{S}_2(t,f)$ respectively.

The assumption that is imposed in the aforementioned framework of using NMF in source separation is that, the trained basis matrix for each source is a sufficient representative for the training data for each source. Some obvious drawbacks of this assumption are that the number of bases can not be determined analytically and the trained matrices do not capture the dynamic information for the source signals. In addition, NMF may cause high overlap among sources due to accepting the whole span of the bases as representations.

In this work, the initial estimated $\widetilde{\boldsymbol{S}}_1$ and $\widetilde{\boldsymbol{S}}_2$ in (10) are treated as distorted 2D signals (images) that need to be restored. MMSE estimation is used as a post process to find better estimates for the source signals.

4. MMSE estimation for post enhancement

We first need to build models for the correct/expected spectrogram frames that the sources $\widetilde{\boldsymbol{S}}_1$ and $\widetilde{\boldsymbol{S}}_2$ should have. For example, the sequence of PSD frames in the spectrogram $\boldsymbol{S}_1^{train}$ in equation (8) can be seen as valid PSD frames that the spectrogram of the first source can have. The training signal spectrograms $\boldsymbol{S}_1^{train}$ and $\boldsymbol{S}_2^{train}$ can be used to train Gaussian mixture models GMM_1 and GMM_2 for the valid PSD frames that can be seen in each source respectively. Then, how far the statistics of the spectrograms $\widetilde{\boldsymbol{S}}_1$ and $\widetilde{\boldsymbol{S}}_2$ from the trained GMM_1 and GMM_2 respectively are learned which are considered as the measurements of the amount of distortions that exist in the spectrograms $\widetilde{\boldsymbol{S}}_1$ and $\widetilde{\boldsymbol{S}}_2$. Based on the amount of the existed distortions and the GMMs that model the valid frames, MMSE estimates are used to find a better solution for each source spectrogram $\widetilde{\boldsymbol{S}}_1$ and $\widetilde{\boldsymbol{S}}_2$. To consider the dynamic information of the source signals, we deal with multiple PSD frames stacked together in one column for training the GMMs and for the MMSE estimates in the enhancement stage. To avoid dealing with the gain differences between the training and separated signals, we normalize each column (stacked PSD frames) using the ℓ^2 norm. To avoid dealing with the nonnegativity constraints we enhance the signals in the log-spectrogram domain. The overall idea of post enhancement here can be seen as a shape or pattern correction. The patterns that exist in the training data spectrograms are used to enhance the NMF separated signal spectrograms through the MMSE estimates.

4.1. Training the source GMMs

First, we stack L frames of the training data spectrogram $\boldsymbol{S}^{train}$ for a given source into one super-frame as in [9, 10, 11]. Each super-frame is normalized and its logarithm is calculated. We form a super-matrix with columns containing the logarithm of the normalized super-frames as shown in Figure 1. We pass a

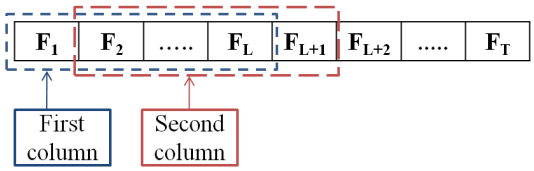

Figure 1: Columns construction and sliding windows with length L frames.

window with length L frames on the training data spectrogram $\boldsymbol{S}_i^{train}$ to select the first column of the super-matrix, then we shift or slide the window by one frame to choose the next super-frame. The super-frames for each source are used to train a GMM. The GMM for a random vector $\boldsymbol{x}$ is defined as

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \frac{\pi_k}{(2\pi)^{d/2} \left|\boldsymbol{\Sigma}_k\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_k\right)\right\}, \tag{14}$$

where K is the number of Gaussian mixture components, π_k is the mixture weight, d is the vector dimension, $\boldsymbol{\mu}_k$ is the mean vector and $\boldsymbol{\Sigma}_k$ is the diagonal covariance matrix of the k^{th} Gaussian model. In training the GMM, the expectation maximization (EM) algorithm [4] is used to learn the GMM parameters $(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \ \forall k = \{1, 2, ..., K\})$ for each source given the logarithm of its normalized super-frames as training data. After training the GMM parameters using each source training data, we will have trained GMM_1 for the first source and GMM_2 for the second source.

4.2. Learning the distortion

We need to learn how much the spectrogram $\widetilde{\boldsymbol{S}}$ for a given source in (10) is distorted compared with its corresponding trained GMM. First, we need to form a super-matrix for each $\widetilde{\boldsymbol{S}}$ in (10). We attach $L-1$ frames with values close to zeros to the far left and right to each spectrogram $\widetilde{\boldsymbol{S}}$. Then we start forming super-frames with L stacked frames for the spectrogram $\widetilde{\boldsymbol{S}}$ as we did during training the GMMs in Section 4.1. Every super-frame is normalized and its logarithm is calculated and used to form a super-matrix $\boldsymbol{Q}$ for its corresponding spectrogram $\widetilde{\boldsymbol{S}}$. The normalization values for the super-frames are saved to be used later. Data corresponding to each PSD frame in $\widetilde{\boldsymbol{S}}$ will appear L times in its corresponding super-matrix $\boldsymbol{Q}$ as sub-vectors in the corresponding super-frame columns. Each column $\boldsymbol{q}_n$ in $\boldsymbol{Q}$ can be seen as a clean observation $\boldsymbol{x}_n$ with additive noise $\boldsymbol{e}$ as follows:

$$\boldsymbol{q}_n = \boldsymbol{x}_n + \boldsymbol{e}, \tag{15}$$

where $\boldsymbol{x}_n$ is the unknown desired pattern that corresponds to the observation $\boldsymbol{q}_n$ and needs to be estimated under a trained GMM from section 4.1, $\boldsymbol{e}$ is the logarithm of a distortion operator, which is modeled here by a Gaussian distribution with zero mean and diagonal covariance matrix $\boldsymbol{\Psi}$ as $\mathcal{N}(\boldsymbol{e}|\boldsymbol{0}, \boldsymbol{\Psi})$. The uncertainty $\boldsymbol{\Psi}$ is trained directly from all columns $\boldsymbol{q} = \{\boldsymbol{q}_1, .., \boldsymbol{q}_n, .., \boldsymbol{q}_N\}$ in $\boldsymbol{Q}$, where N is the number of columns in the matrix $\boldsymbol{Q}$. The uncertainty $\boldsymbol{\Psi}$ can be iteratively learned using the expectation maximization (EM) algorithm. Given the GMM parameters which are considered fixed here, the update of $\boldsymbol{\Psi}$ is found based on the sufficient statistics $\hat{\boldsymbol{z}}_n$ and $\hat{R}_n$ as follows [12, 13, 14, 15]:

$$\boldsymbol{\Psi} = \text{diag}\left\{\frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{q}_n\boldsymbol{q}_n^T - \boldsymbol{q}_n\hat{\boldsymbol{z}}_n^T - \hat{\boldsymbol{z}}_n\boldsymbol{q}_n^T + \hat{\boldsymbol{R}}_n\right)\right\}, \tag{16}$$

where the "diag" operator sets all the off-diagonal elements of a matrix to zero, N is the number of columns in matrix $\boldsymbol{Q}$, and the sufficient statistics $\hat{\boldsymbol{z}}_n$ and $\hat{R}_n$ can be updated using $\boldsymbol{\Psi}$ from the previous iteration as follows:

$$\hat{\boldsymbol{z}}_n = \sum_{k=1}^{K}\gamma_{kn}\hat{\boldsymbol{z}}_{kn}, \quad \text{and} \quad \hat{\boldsymbol{R}}_n = \sum_{k=1}^{K}\gamma_{kn}\hat{\boldsymbol{R}}_{kn}, \tag{17}$$

where

$$\gamma_{kn} = \left[\frac{\pi_k\mathcal{N}(\boldsymbol{q}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k + \boldsymbol{\Psi})}{\sum_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{q}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j + \boldsymbol{\Psi})}\right], \tag{18}$$

$$\hat{\boldsymbol{R}}_{kn} = \boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_k\left(\boldsymbol{\Sigma}_k + \boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Sigma}_k^T + \hat{\boldsymbol{z}}_{kn}\hat{\boldsymbol{z}}_{kn}^T, \tag{19}$$

and

$$\hat{\boldsymbol{z}}_{kn} = \boldsymbol{\mu}_k + \boldsymbol{\Sigma}_k\left(\boldsymbol{\Sigma}_k + \boldsymbol{\Psi}\right)^{-1}\left(\boldsymbol{q}_n - \boldsymbol{\mu}_k\right). \tag{20}$$

$\boldsymbol{\Psi}$ is considered as a general uncertainty measurement over all the observations in matrix $\boldsymbol{Q}$. $\boldsymbol{\Psi}$ can be seen as a model that summarizes the deformation that exists in all columns in the super-matrix $\boldsymbol{Q}$. Given the trained GMM_1, GMM_2, the super matrices $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ that are corresponding to the distorted spectrograms $\widetilde{\boldsymbol{S}}_1$ and $\widetilde{\boldsymbol{S}}_2$, the uncertainties $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ for the first and second source are calculated iteratively using equations (16) to (20).

4.3. Calculating MMSE estimates

Given the GMM parameters and the uncertainty measurement $\boldsymbol{\Psi}$ for a given source signal, the MMSE estimate of each pattern $\boldsymbol{x}_n$ given its observation $\boldsymbol{q}_n$ under the observation model in equation (15) can be found similar to [12, 13, 14, 15] as follows:

$$\hat{\boldsymbol{x}}_n = \sum_{k=1}^{K}\gamma_{kn}\left[\boldsymbol{\mu}_k + \boldsymbol{\Sigma}_k\left(\boldsymbol{\Sigma}_k + \boldsymbol{\Psi}\right)^{-1}\left(\boldsymbol{q}_n - \boldsymbol{\mu}_k\right)\right], \tag{21}$$

where

$$\gamma_{kn} = \left[\frac{\pi_k\mathcal{N}(\boldsymbol{q}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k + \boldsymbol{\Psi})}{\sum_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{q}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j + \boldsymbol{\Psi})}\right]. \tag{22}$$

The distortion measurement $\boldsymbol{\Psi}$ in the term $\boldsymbol{\Sigma}_k\left(\boldsymbol{\Sigma}_k + \boldsymbol{\Psi}\right)^{-1}$ in equation (21) plays an important role in this framework. When the uncertainty $\boldsymbol{\Psi}$ of the observations $\boldsymbol{q}$ for a given source is high, the MMSE estimate of $\boldsymbol{x}$, relies more on the trained GMM of $\boldsymbol{x}$. When the uncertainty of the observations $\boldsymbol{q}$ is low, the MMSE estimate of $\boldsymbol{x}$, relies more on the observation $\boldsymbol{q}$.

The model in equation (15) expresses the normalized super-columns before calculating the logarithm of the spectrogram $\widetilde{\boldsymbol{S}}$ as a distorted image with a multiplicative deformation diagonal

matrix. For the normalized super-frame columns $\frac{\boldsymbol{s}_n}{\|\boldsymbol{s}_n\|_2}$ of $\widetilde{\boldsymbol{S}}$ there is a deformation matrix $\boldsymbol{E}$ with log-normal distribution that is applied to the correct pattern that we need to estimate $\hat{\boldsymbol{s}}_n$ as follows:

$$\frac{\boldsymbol{s}_n}{\|\boldsymbol{s}_n\|_2} = \boldsymbol{E}\hat{\boldsymbol{s}}_n. \tag{23}$$

The uncertainty for $\boldsymbol{E}$ is represented in the covariance matrix $\boldsymbol{\Psi}$. The MMSE estimation based post enhancement here can be seen as performing denoising under multiplicative noise. We believe this is beneficial since the additive noise is assumed to be removed by NMF.

After calculating $\hat{\boldsymbol{x}}_n, \forall n \in \{1, .., N\}$ we calculate the exponent for each entry of $\hat{\boldsymbol{x}}_n, \forall n \in \{1, .., N\}$ and form a matrix $\boldsymbol{R}$. The procedures in sections 4.2 and 4.3 are repeated for each source to get $\boldsymbol{R}_1$ for the first source and $\boldsymbol{R}_2$ for the second source. The norm for each super-columns that were calculated in section 4.2 are used to scale their corresponding super-columns in $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$. The columns of $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are scaled by multiplying each super-frame (column) with its corresponding norm from section 4.2. The norm rescaling is used to preserve the gain differences between the two source signals. We convert the scaled super-frames of $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ into the original size of the spectrograms by reframing their super-frames. Since every PSD frame appears L times in different L consequent super-frames, we take the average to find the final enhanced spectrograms $\overline{\boldsymbol{S}}_1$ and $\overline{\boldsymbol{S}}_2$. The spectrograms $\overline{\boldsymbol{S}}_1$ and $\overline{\boldsymbol{S}}_2$ are then used in the Wiener filters $\overline{\boldsymbol{H}}_1$ and $\overline{\boldsymbol{H}}_2$ to find the final source STFTs as follows:

$$\overline{\boldsymbol{H}}_1 = \frac{\overline{\boldsymbol{S}}_1}{\overline{\boldsymbol{S}}_1 + \overline{\boldsymbol{S}}_2}, \qquad \overline{\boldsymbol{H}}_2 = \frac{\overline{\boldsymbol{S}}_2}{\overline{\boldsymbol{S}}_1 + \overline{\boldsymbol{S}}_2}, \tag{24}$$

$$\hat{\overline{S}}_1(t,f) = \overline{\boldsymbol{H}}_1(t,f)\, Y(t,f), \quad \hat{\overline{S}}_2(t,f) = \overline{\boldsymbol{H}}_2(t,f)\, Y(t,f), \tag{25}$$

where the divisions are done element-wise which is similar to equations (11) and (12). The use of the Wiener filters here is the only way to guarantee that the two estimated source spectrograms add up to the mixed signal spectrogram. The estimated source signals $\hat{\overline{s}}_1(t)$ and $\hat{\overline{s}}_2(t)$ can be found by using inverse STFT of $\hat{\overline{S}}_1(t,f)$ and $\hat{\overline{S}}_2(t,f)$ respectively.

5. Experiments and Discussion

We applied the proposed algorithm to separate a speech signal from a background piano music signal. Our main goal was to get a clean speech signal from a mixture of speech and piano signals. We simulated our algorithm on a collection of speech and piano data at 16kHz sampling rate. For speech data, we used the training and testing male speech data from the TIMIT database. For music data, we downloaded piano music data from the piano society web site [16]. We used 12 pieces with approximate 50 minutes total duration from different composers but from a single artist for training and left out one piece for testing. The PSD for the speech and music data were calculated by using the STFT: A Hamming window with 480 points length and 60% overlap was used and the FFT was taken at 512 points, the first 257 FFT points only were used since the conjugate of the remaining 255 points are involved in the first points. We trained 128 basis vectors for each source, which makes the size of $\boldsymbol{B}_{\text{speech}}$ and $\boldsymbol{B}_{\text{music}}$ matrices to be 257×128. The mixed data was formed by adding random portions of the test music file to 20 speech files from the test data of the TIMIT database at different speech-to-music ratio (SMR) values in dB. The audio power levels of each file were found using the "speech voltmeter" program from the G.191 ITU-T STL software suite [17]. For each SMR value, we obtained 20 mixed utterances this way.

Performance measurements of the separation algorithm were done using the signal to distortion ratio (SDR) and the signal to interference ratio (SIR) [18]. The average SDR and SIR over the 20 test utterances are reported. The source to distortion ratio (SDR) is defined as the ratio of the target energy to all errors in the reconstructed signal. The target signal is defined as the projection of the predicted signal onto the original speech signal. Signal to interference ratio (SIR) is defined as the ratio of the target energy to the interference error due to the music signal only. The higher SDR and SIR we measure the better performance we achieve.

Table 1 shows SDR and SIR of the separated speech signal using NMF without post enhancement and NMF with post enhancement using MMSE estimates with different values of GMM components K and the number of the stacked frames L. The second column of the table, shows the separation results of using just NMF with spectral masks without post enhancement as shown in equations (12) and (13). The third and fourth columns show the results of using NMF with MMSE estimation based post enhancement with the Wiener filters as shown in equations (24) and (25). The choice for K and L was done by trying different combinations. In this work, we chose the same value for L for both sources and also for K. The shown results are just examples for the improvements that can be achieved. Better results can be achieved for different combinations of K and L. Suitable values for K and L can be found using validation data.

Table 1: SDR and SIR in dB for the estimated speech signal.

SMR	NMF		NMF+Post MMSE			
			$L=5, K=128$		$L=3, K=32$	
dB	SDR	SIR	SDR	SIR	SDR	SIR
-5	1.79	5.01	**3.99**	**9.79**	2.96	6.88
0	4.51	8.41	**6.13**	**12.03**	5.61	9.89
5	7.99	12.36	**9.18**	**15.49**	9.00	13.72
10	10.30	16.48	**11.31**	**18.62**	11.29	17.34
15	12.00	20.05	12.66	**21.16**	**13.05**	20.48
20	13.07	24.93	13.95	**25.67**	**14.32**	25.16

As we can see from the table, the proposed NMF with post enhancement using MMSE estimates improves the separation performance comparing with just using NMF. Increasing the value of L improves the performance especially at low SMR values but it requires increasing the value of K. The best choice for K usually depends on the nature, and the size of the training data, and also on the value of L. It is important to note that, applying MMSE estimates directly on the mixed signal without using NMF (not shown in the table) gives worse results than just using NMF.

6. Conclusion

In this work, we improved the quality of NMF based source separation by employing a novel MMSE estimation technique based on trained GMMs. The distortion was learned online from the NMF separated signal spectrograms. The dynamics or the sequential information of the sources were considered by enhancing multiple frames of the spectrograms at once. The results show that, the proposed MMSE estimation based post enhancement improves the quality of the NMF separated sources.

7. References

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[2] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *International Conference on Spoken Language Processing (INTERSPEECH)*, 2006.

[3] E. M. Grais and H. Erdogan, "Spectro-temporal post-smoothing in NMF based single-channel source separation," in *European Signal Processing Conference (EUSIPCO)*, 2012.

[4] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, 1977.

[5] C. Fevotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[6] E. M. Grais and H. Erdogan, "Hidden Markov Models as priors for regularized nonnegative matrix factorization in single-channel source separation," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2012.

[7] ——, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," in *International Conference on Digital Signal Processing*, 2011.

[8] ——, "Regularized nonnegative matrix factorization using gaussian mixture priors for supervised single channel source separation," *Computer Speech and Language*, vol. 27, no. 3, pp. 746–762, May 2013.

[9] ——, "Single channel speech music separation using nonnegative matrix factorization with sliding window and spectral masks," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011.

[10] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, 2011.

[11] J. F. Gemmeke, T. Virtanen, and Y. Sun, "Noise robust exemplar-based connected digital recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

[12] URL, "http://arxiv.org/abs/1302.7283," 2013.

[13] A.-V. Rosti and M. Gales, "Generalised linear gaussian models," CUED/F-INFENG/TR.420, University of Cambridge, Tech. Rep., 2001.

[14] A.-V. I. Rosti and M. J. F. Gales, "Factor analysed hidden markov models for speech recognition," *Computer Speech and Language, Issue 2*, vol. 18, pp. 181–200, 2004.

[15] Z. Ghahramani and G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," CRG-TR-96-1, University of Toronto, Canada, Tech. Rep., Feb. 1997.

[16] URL, "http://pianosociety.com," 2009.

[17] ——, "http://www.itu.int/rec/T-REC-G.191/en," 2009.

[18] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–69, Jul. 2006.