Adaptation of speaker-specific bases in non-negative matrix factorization for single channel speech-music separation

Emad M. Grais and Hakan Erdogan

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

Abstract

This paper introduces a speaker adaptation algorithm for non-negative matrix factorization (NMF) models. The proposed adaptation algorithm is a combination of Bayesian and subspace model adaptation. The adapted model is used to separate speech signal from a background music signal in a single record. Training speech data for multiple speakers is used with NMF to train a set of basis vectors as a general model for speech signals. The probabilistic interpretation of NMF is used to achieve Bayesian adaptation to adjust the general model with respect to the actual properties of the speech signals that is observed in the mixed signal. The Bayesian adapted model is adapted again by a linear transform, which changes the subspace that the Bayesian adapted model spans to better match the speech signal that is in the mixed signal. The experimental results show that combining Bayesian with linear transform adaptation improves the separation results.

Index Terms: Model adaptation, single channel source separation, source separation, speech music separation, and nonnegative matrix factorization.

1. Introduction

Model adaptation is usually an alternative approach that is used to overcome the problem of the lack of enough training data to accurately model the actual characteristics of any signal. A general model is built first from general training data, then this model is adapted to capture the properties of the target data.

In speech recognition, adaptation is used intensively to adapt the parameters of the speech models [1]. Model adaptation is also used in single channel source separation applications to adapt the source signal models to better represent the actual properties of the observed signals in the mixed signal. In [2], Bayesian adaptation was used to adapt the GMM model for each source signal. The data that was used to adapt the models is estimated from the observed mixed signal directly. In [3, 4], the adaptation of the set of basis vectors that models every source signal was introduced, and the adaptation is done within the separation process without any need for an extra adaptation stage.

Most algorithms that use NMF to separate source signals from a mixture of source signals assume that, there is enough training data available for each source. NMF uses these data in magnitude spectral domain to train a set of basis vectors for each source. These sets of bases are used with NMF to estimate the source signals from the mixture. This kind of algorithms produce good results when enough training data is available, and the spectral characteristics of the training data are similar to those of the data in the mixture. In speech-music separation, sometimes finding enough training speech data for a specific speaker that is in the mixture signal is not easy. Building a source model using little training data leads to a poor model that is incapable of capturing the actual characteristics of the source signal. Also using other speakers speech signals that are not in the mixture as a training data leads to mismatch between training and target data, which decreases the quality of the obtained solution. The key idea in this paper is, rather than using the small training data for a specific speaker to train a set of basis vectors with more entries to estimate, we train a general set of basis vectors using enough speech signals from multiple speakers. Then we adapt these basis vectors to better match the target data. First, we use NMF and training speech data of many speakers to train a general set of basis vectors. Second, we adapt these bases using a small amount of training data of a specific speaker to get speaker-specific bases. The adapted bases are used to separate the speech signal of the same speaker from the background music. Here we assume that there is a small amount of isolated training speech signal of the speaker that is in the mixture signal. To use the adaptation data optimally we propose to adapt the general model twice. First, we adapt the general model using Bayesian adaptation which relies on the probabilistic interpretation of the standard NMF that is represented in [5, 6]. Second, the resulting adapted model is adapted again by a linear transformation similar to MLLR [7]. This linear transform is found by using NMF and the adaptation data. The novelty of this work is in combining the Bayesian and linear regression adaptations to adapt a set of speech basis vectors, and also in introducing the update rules for the multiplicative adaptation matrix.

The remainder of this paper is organized as follows: Section 2 shows a mathematical description of the source separation problem. In section 3, a brief explanation about NMF and how we use it to train the basis vectors for each source is given. In section 4, the two adaptation algorithms are proposed and bases adaptation is explained. Section 5 shows the separation process. In the remaining sections, we present our observations and the results of our experiments.

2. Problem formulation

Given an observed mixed signal $x(t)$ which is a mixture of speech $s(t)$ and music signals $m(t)$, the single channel source separation techniques aim to find estimates for $s(t)$ and $m(t)$ from $x(t)$. We propose to solve this problem in the short time Fourier transform (STFT) domain. Let $X(t, f)$ be the STFT of $x(t)$, where t represents the frame index and f is the frequency-index. Due to the linearity of the STFT, we have:

$$X(t, f) = S(t, f) + M(t, f), \tag{1}$$

$$|X(t,f)|\, e^{j\phi_X(t,f)} = |S(t,f)|\, e^{j\phi_S(t,f)} + |M(t,f)|\, e^{j\phi_M(t,f)}. \quad (2)$$

In this work, it is assumed that all phase angles are the same, that is $\phi_S(t,f) = \phi_M(t,f) = \phi_X(t,f)$. Hence, we can write the magnitude spectrogram of the measured signal as the sum of source signals' magnitude spectrograms.

$$\boldsymbol{X} = \boldsymbol{S} + \boldsymbol{M}.^{1} \quad (3)$$

$\boldsymbol{S}$ and $\boldsymbol{M}$ are unknown magnitude spectrograms, and need to be estimated using observed data and training speech and music spectra. The magnitude spectrogram for the observed signal $x(t)$ is obtained by taking the magnitude of the DFT of the windowed signal.

3. Non-negative matrix factorization

Non-negative matrix factorization is used to decompose any nonnegative matrix $\boldsymbol{V}$ into a nonnegative basis vectors matrix $\boldsymbol{B}$ and a nonnegative weights matrix $\boldsymbol{W}$.

$$\boldsymbol{V} \approx \boldsymbol{BW}. \quad (4)$$

The matrices $\boldsymbol{B}$ and $\boldsymbol{W}$ can be found by solving the divergence cost function [8] which is preferred to be used in audio source separation applications [9]. The divergence cost function yields the following optimization problem:

$$\min_{\boldsymbol{B},\boldsymbol{W}} D\left(\boldsymbol{V} \,||\, \boldsymbol{BW}\right), \quad (5)$$

where

$$D\left(\boldsymbol{V} \,||\, \boldsymbol{BW}\right) = \sum_{i,j} \left(\boldsymbol{V}_{i,j} \log \frac{\boldsymbol{V}_{i,j}}{(\boldsymbol{BW})_{i,j}} - \boldsymbol{V}_{i,j} + (\boldsymbol{BW})_{i,j} \right),$$

subject to elements of $\boldsymbol{B}, \boldsymbol{W} \geq 0$. The solution for equation (5) can be computed by alternating updates of $\boldsymbol{B}$ and $\boldsymbol{W}$ as follows:

$$\boldsymbol{B} \leftarrow \boldsymbol{B} \otimes \frac{\dfrac{\boldsymbol{V}}{\boldsymbol{BW}} \boldsymbol{W}^T}{\mathbf{1}\boldsymbol{W}^T}, \quad (6)$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \frac{\boldsymbol{B}^T \dfrac{\boldsymbol{V}}{\boldsymbol{BW}}}{\boldsymbol{B}^T\mathbf{1}}, \quad (7)$$

where $\mathbf{1}$ is a matrix of ones with the same size of $\boldsymbol{V}$, the operations $\otimes$ and all divisions are element wise multiplication and division respectively.

3.1. Probabilistic perspective of NMF

As shown in [5, 6], each entry $v_{k,j}$ of the matrix $\boldsymbol{V}$ can be modelled by Poisson distribution as follows:

$$p(v_{k,j}|b_{k,1:I}, w_{1:I,j}) = PO(v_{k,j}; \sum_{i}^{I} b_{k,i} w_{i,j}), \quad (8)$$

where $b_{k,1:I}$ denotes the k^{th} column of $\boldsymbol{B}$, $w_{1:I,j}$ the j^{th} row of $\boldsymbol{W}$, respectively, and the Poisson distribution defined as

$$PO(v;\lambda) = \frac{e^{-\lambda}\lambda^v}{\Gamma(v+1)}, \quad (9)$$

[1] The notations here are as follows: bold capital letters are for matrices, bold small letters are for vectors others are for scalars.

where $\Gamma(v)$ is the gamma function. Assuming that each entry $v_{k,j}$ is statistically independent conditional on $\boldsymbol{B}$ and $\boldsymbol{W}$, the model can be denoted by:

$$p(\boldsymbol{V}|\boldsymbol{B},\boldsymbol{W}) = \prod_{k,j} \frac{e^{-[\boldsymbol{BW}]_{k,j}} [\boldsymbol{BW}]_{k,j}^{[\boldsymbol{V}]_{k,j}}}{\Gamma([\boldsymbol{V}]_{k,j}+1)}. \quad (10)$$

The maximum likelihood solution is found by

$$(\boldsymbol{B},\boldsymbol{W}) = \arg\max_{\boldsymbol{B},\boldsymbol{W}} \log p(\boldsymbol{V}|\boldsymbol{B},\boldsymbol{W}), \quad (11)$$

where

$$\log p(\boldsymbol{V}|\boldsymbol{B},\boldsymbol{W}) =$$

$$\sum_{k,j} -[\boldsymbol{BW}]_{k,j} + [\boldsymbol{V}]_{k,j} \log([\boldsymbol{BW}]_{k,j}) - \log(\Gamma([\boldsymbol{V}]_{k,j}+1)).$$

We can see that finding the maximum likelihood solution is equivalent to solving the objective function (5). The advantage of using NMF in probabilistic framework is the ability to put priors on every entry of the matrices $\boldsymbol{B}$ and $\boldsymbol{W}$ [6]. In this work, we will use the advantage of putting priors only on the entries of the bases matrix $\boldsymbol{B}$ as we will show in next sections.

3.2. Basis vectors matrix prior $p(\boldsymbol{B})$

In [6], the prior on each basis vector matrix entry is assumed to be independently drawn from a Gamma distribution:

$$p(b_{k,i}) = g(b_{k,i}; \alpha_{k,i}, \beta_{k,i}^{-1}) = \frac{b_{k,i}^{\alpha_{k,i}-1} \beta_{k,i}^{\alpha_{k,i}} e^{-b_{k,i}\beta_{k,i}}}{\Gamma(\alpha_{k,i})}. \quad (12)$$

The hyperparameters $\alpha_{k,i}$ and $\beta_{k,i}$ of the model can be selected individually for each bases matrix entry. It is also assumed that $p(\boldsymbol{B}) = \prod_{i=1}^{I}\prod_{k=1}^{F} p(b_{k,i})$ then we have

$$\log p(\boldsymbol{B}) =^{+} \sum_{i=1}^{I}\sum_{k=1}^{F} (\alpha_{k,i}-1)\log(b_{k,i}) - b_{k,i}\beta_{k,i}. \quad (13)$$

Here $=^{+}$ denotes equal up to irrelevant constant terms (i.e. $p \propto q \iff \log p =^{+} \log q$). The joint posterior distribution is given by Bayes rule $p(\boldsymbol{B},\boldsymbol{W}|\boldsymbol{V}) \propto p(\boldsymbol{V}|\boldsymbol{B},\boldsymbol{W})P(\boldsymbol{B},\boldsymbol{W})$ which factorises to $p(\boldsymbol{V}|\boldsymbol{B},\boldsymbol{W})P(\boldsymbol{B})P(\boldsymbol{W})$. The MAP estimate can be found as

$$\arg\max_{\boldsymbol{B},\boldsymbol{W}} \left[\log p(\boldsymbol{V}|\boldsymbol{B},\boldsymbol{W}) + \log p(\boldsymbol{W}) + \log p(\boldsymbol{B})\right]. \quad (14)$$

In this work, we do not use prior on the gain matrix $p(\boldsymbol{W})$. We substitute the terms in (14) with the equation (11) and (13). The MAP estimator can be derived [6], and the update rules for each element in the bases matrix $\boldsymbol{B}$ and the gain matrix $\boldsymbol{W}$ is given as

$$b_{k,i} \leftarrow b_{k,i} \frac{\frac{(\alpha_{k,i}-1)}{b_{k,i}} + \sum_{j=1}^{K} w_{i,j} \frac{v_{k,j}}{\sum_{i=1}^{I} b_{k,i} w_{i,j}}}{\beta_{k,i} + \sum_{j'=1}^{K} w_{i,j'}}, \quad (15)$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \frac{\boldsymbol{B}^T \dfrac{\boldsymbol{V}}{\boldsymbol{BW}}}{\boldsymbol{B}^T\mathbf{1}}. \quad (16)$$

Notice that the update rule (15) differs from the basic NMF update (6) only by additive terms in the numerator and denominator, which are due to the priors.

3.3. Training the bases

Given a set of training data for music and speech of multiple speakers signals, The STFT is computed for each signal, and the magnitude spectrogram $\boldsymbol{S}_{\text{train}}$ and $\boldsymbol{M}_{\text{train}}$ of speech and music respectively are calculated. Then NMF is used to decompose these spectrograms into bases and weights matrices as follows:

$$\boldsymbol{S}_{\text{train}} \approx \boldsymbol{B}_{\text{speech}} \boldsymbol{W}_{\text{speech}}. \tag{17}$$

$$\boldsymbol{M}_{\text{train}} \approx \boldsymbol{B}_{\text{music}} \boldsymbol{W}_{\text{music}}. \tag{18}$$

We use the update rules in equations (6) and (7) to solve equations (17) and (18). $\boldsymbol{S}$ and $\boldsymbol{M}$ have normalized columns, and after each iteration, we normalize the columns of $\boldsymbol{B}_{\text{speech}}$ and $\boldsymbol{B}_{\text{music}}$. All the matrices $\boldsymbol{B}$ and $\boldsymbol{W}$ are initialized by positive random noise. We call the trained bases matrix $\boldsymbol{B}_{\text{speech}}$ a general model for multiple speakers speech signals, and this matrix needs to be adapted to specific speaker speech signals.

4. Speech model adaptation

Given the general speech model which represents multiple speakers speech signals $\boldsymbol{B}_{\text{speech}}$, the goal now is to adapt this model using a small amount of a specific-speaker speech signals to better match the target speech signal that is in the mixture. We combine two adaptation techniques to adapt the speech model. First adaptation algorithm is the Bayesian adaption, which is driven from the probabilistic framework of NMF in equation (15). Second adaptation algorithm which we introduce in this paper is driven from linear regression, which aims to change the subspace of the model to better match the target data.

4.1. Speech bases adaptation

In this work, we assume that we have a small adaptation data of speaker-specific speech signal s_{adapt}. The goal now is to use the spectrogram of this new data $\boldsymbol{S}_{\text{adapt}}$ to adapt the general bases matrix $\boldsymbol{B}_{\text{speech}}$ to become speaker specific bases matrix $\boldsymbol{B}_s$. We will use first the Bayesian adaptation in equation (15) by replacing β^{-1} values with the entries of $\boldsymbol{B}_{\text{speech}}$, and $\alpha = 2$ everywhere inspired from [6]. This choice makes the mode of the Gamma distribution equal to the general model bases $\boldsymbol{B}_{\text{speech}}$. Which means that the general model is used as a prior for $\boldsymbol{B}_s$, so the update rules will be as follows:

$$\boldsymbol{B}_s \leftarrow \boldsymbol{B}_s \otimes \frac{\frac{\boldsymbol{1}'}{\boldsymbol{B}_s} + \frac{\boldsymbol{S}_{\text{adapt}}}{\boldsymbol{B}_s \boldsymbol{W}_a} \boldsymbol{W}_a^T}{\frac{\boldsymbol{1}'}{\boldsymbol{B}_{\text{speech}}} + \boldsymbol{1} \boldsymbol{W}_a^T}, \tag{19}$$

$$\boldsymbol{W}_a \leftarrow \boldsymbol{W}_a \otimes \frac{\boldsymbol{B}_s^T \frac{\boldsymbol{S}_{\text{adapt}}}{\boldsymbol{B}_s \boldsymbol{W}_a}}{\boldsymbol{B}_s^T \boldsymbol{1}}. \tag{20}$$

The matrix $\boldsymbol{B}_s$ is initialized with $\boldsymbol{B}_{\text{speech}}$ and $\boldsymbol{W}_a$ is initialized by positive random noise. Here every division is element wise and $\boldsymbol{1}'$ is a matrix of ones of the same size of $\boldsymbol{B}_{\text{speech}}$.

To use the adaptation data optimally, we use an extra step to adapt the bases matrix $\boldsymbol{B}_s$ by multiplying it with an adaptation matrix $\boldsymbol{A}$. The final user specific bases matrix is found as $\boldsymbol{B}_{sf} = \boldsymbol{A}\boldsymbol{B}_s$, $\boldsymbol{A}$ is the adaptation matrix which is unknown and needs to be calculated as follows:

$$D\left(\boldsymbol{S}_{\text{adapt}} \| \boldsymbol{B}_{sf} \boldsymbol{W}_{a2}\right) = D\left(\boldsymbol{S}_{\text{adapt}} \| \left(\boldsymbol{A}\boldsymbol{B}_s\right) \boldsymbol{W}_{a2}\right), \tag{21}$$

$$\boldsymbol{A}, \boldsymbol{W}_{a2} = \arg \min_{\boldsymbol{A}, \boldsymbol{W}_{a2}} D(\boldsymbol{S}_{\text{adapt}} \| \boldsymbol{A}\boldsymbol{B}_s \boldsymbol{W}_{a2}). \tag{22}$$

We employ alternating minimization for equation (22) by fixing $\boldsymbol{B}_s \boldsymbol{W}_{a2}$ as one matrix and first update $\boldsymbol{A}$ using equation (6) as follows:

$$\boldsymbol{A} \leftarrow \boldsymbol{A} \otimes \frac{\frac{\boldsymbol{S}_{\text{adapt}}}{\boldsymbol{A}(\boldsymbol{B}_s \boldsymbol{W}_{a2})} \left(\boldsymbol{B}_s \boldsymbol{W}_{a2}\right)^T}{\boldsymbol{1} \left(\boldsymbol{B}_s \boldsymbol{W}_{a2}\right)^T}, \tag{23}$$

then we fix $\boldsymbol{A}\boldsymbol{B}_s$ as one matrix and find $\boldsymbol{W}_{a2}$ using equation (7) as follows:

$$\boldsymbol{W}_{a2} \leftarrow \boldsymbol{W}_{a2} \otimes \frac{\left(\boldsymbol{A}\boldsymbol{B}_s\right)^T \frac{\boldsymbol{S}_{\text{adapt}}}{\boldsymbol{A}\boldsymbol{B}_s \boldsymbol{W}_{a2}}}{\left(\boldsymbol{A}\boldsymbol{B}_s\right)^T \boldsymbol{1}}, \tag{24}$$

$\boldsymbol{B}_s$ always fixed in both equations. We need only to use $\boldsymbol{A}$ to find the final adapted bases matrix as

$$\boldsymbol{B}_{sf} = \boldsymbol{A}\boldsymbol{B}_s. \tag{25}$$

Since we assume that, the adaptation data is small then it is better if there are fewer values to be estimated in the matrix $\boldsymbol{A}$. We enforce the adaptation matrix $\boldsymbol{A}$ to be diagonal with extra nonzero column by initializing it this way since the update rule for $\boldsymbol{A}$ in equation (23) is element-wise multiplication. We also add an extra row in matrix $\boldsymbol{B}_s$ with ones to enable a bias term similar to MLLR. By multiplying the adaptation matrix $\boldsymbol{A}$ with $\boldsymbol{B}_s$ the columns of the adapted matrix $\boldsymbol{B}_{sf}$ can span any other subspaces that the adaptation data may lie on which the columns of $\boldsymbol{B}_s$ can not span. We achieved that by estimating fewer parameters for the matrix $\boldsymbol{A}$ rather than using speaker specific data to train the bases matrix with more parameters, especially since the speaker specific training data "adaptation data" is small. After finding the bases matrix $\boldsymbol{B}_{sf}$ which is close to be a speaker specific bases matrix, we use it to separate speech signal of the same speaker from the background music signal.

5. Signal separation and reconstruction

After observing the mixed signal $x(t)$, the magnitude spectrogram $\boldsymbol{X}$ of the mixed signal is computed using STFT. NMF is used to decompose the magnitude spectrogram $\boldsymbol{X}$ of the mixed signal as a linear combination with the trained basis vectors in $\boldsymbol{B}_{sf}$ and $\boldsymbol{B}_{\text{music}}$ as follows:

$$\boldsymbol{X} \approx \left[\boldsymbol{B}_{sf} \;\; \boldsymbol{B}_{\text{music}}\right] \boldsymbol{W}, \tag{26}$$

where $\boldsymbol{B}_{sf}$ and $\boldsymbol{B}_{\text{music}}$ are obtained from equations (25) and (18). Here only the update rule in equation (7) is used to solve equation (26), and the bases matrix is fixed. $\boldsymbol{W}$ is initialized by positive random noise. The initial spectrograms estimate for speech and music signals are respectively calculated as follows: $\tilde{\boldsymbol{S}} = \boldsymbol{B}_{sf} \boldsymbol{W}_S$ and $\tilde{\boldsymbol{M}} = \boldsymbol{B}_{\text{music}} \boldsymbol{W}_M$. Where $\boldsymbol{W}_S$ and $\boldsymbol{W}_M$ are submatrices in matrix $\boldsymbol{W}$ that correspond to the speech and music components respectively in equation (26). The final estimate of the speech signal spectrogram is found as follows:

$$\hat{\boldsymbol{S}} = \boldsymbol{H} \otimes \boldsymbol{X}, \tag{27}$$

where $\otimes$ is element-wise multiplication, and $\boldsymbol{H}$ is the Wiener filter which is defined as follows:

$$\boldsymbol{H} = \frac{\tilde{\boldsymbol{S}}^2}{\tilde{\boldsymbol{S}}^2 + \tilde{\boldsymbol{M}}^2}. \tag{28}$$

Where $(.)^2$ and division are element-wise operations. Wiener filter works here as a soft mask for the observed mixed signal, which scales the magnitude of the mixed signal at every frequency component with values between 0 and 1 to find their corresponding frequency component values in the estimated speech signal. After finding the contribution of the speech signal in the mixed signal, the estimated speech signal $\hat{s}(t)$ can be found by using inverse STFT to the estimated speech spectrogram $\hat{\boldsymbol{S}}$ with the phase angle of the mixed signal.

6. Experiments and Results

We simulated the proposed algorithms on a collection of speech and piano music data at 16kHz sampling rate. For the general training speech data, we used 1000 utterances from multiple male speakers from the Timit database. For testing, we applied the proposed algorithm on 20 different speakers, and we averaged the results. We used 20 utterances from different 20 speakers that are not included in the training data for testing and adaptation. We used around 12 seconds for each speaker as adaptation data to adapt the general bases matrix for each speaker individually, which means we got 20 adapted models, one for each speaker. All the speech signals that were used in our experiments are for male speakers. For music data, we downloaded piano music from piano society web site [10]. We used 38 pieces from different composers but from a single artist for training and left out one piece for the testing stage. The magnitude spectrograms for the training speech and music data are calculated by using the STFT, a Hamming window was used and the FFT was taken at 512 points, the first 257 FFT points only were used since the remaining points are the conjugate of the first 257 points. We trained the general speech bases matrix using 32 basis vectors and the same for the music signal. The test data was formed by adding random portions of the test music file to the 20 speech utterance files at different speech to music ratio (SMR) values in dB. The audio power levels of each file were found using the "audio voltmeter" program from the G.191 ITU-T STL software suite [11]. For each SMR value, we obtained 20 test utterances of different 20 speakers this way. Performance measurement of the separation algorithms was done using signal to noise ratio in the time domain.

We tried to separate the speech signal from the music background using different experiments. In every experiment, we use a different bases matrix for the speech signal. In the first experiment, we tried to separate the mixture using only the general bases matrix $\boldsymbol{B}_{\text{speech}}$ without any adaptation. In the second experiment, we used the adaptation data, which is a speaker specific signal with duration 12 seconds only to train the bases matrix $\boldsymbol{B}_{\text{speaker}}$ from scratch without using the general bases matrix at all. In the third experiment, we used the Bayesian adaptation only to find $\boldsymbol{B}_s$ without the multiplication adaptation. In the fourth experiment, we used the two adaptation algorithms first with Bayesian adaptation to find $\boldsymbol{B}_s$ then we applied the multiplicative adaptation to find $\boldsymbol{B}_{sf}$. Table 1 shows the results of these experiments. These results are the average over 20 different speakers. The results show that using Bayesian adaptation improves the results compared with using the general model directly. Also using multiplication adaptation after Bayesian adaptation improves the results even more than using the Bayesian adaptation only. For the second experiment that uses the small speaker-specific training speech data only to train the bases matrix model without using the general model, we get the worst results in most of SMR except at -5dB case. These results show that if we need to separate a mixture of speech and music signals, and we have a small amount of training speech data of the speaker that is in the mixed signal, the better way to build a speech model is to train a general model using plenty amount of multiple speakers training data, then use the small amount of the speaker specific data to adapt the general model. Audio demonstrations of our experiments are available at http://students.sabanciuniv.edu/grais/speech/assbnmfscsms/

Table 1: Signal to Noise Ratio (SNR) in dB for the separated speech signal for every experiment.

SMR dB	Using only $\boldsymbol{B}_{\text{speech}}$	Using only $\boldsymbol{B}_{\text{speaker}}$	Using only $\boldsymbol{B}_s$	Using $\boldsymbol{B}_{sf}$
-5	3.15	**3.43**	3.17	3.26
0	4.92	4.76	5.26	**5.29**
5	6.32	5.88	6.83	**6.85**
10	7.33	6.43	8.00	**8.05**
15	7.79	6.81	8.57	**8.72**
20	7.97	7.00	8.92	**9.03**

7. CONCLUSION

In this work, we proposed a model adaptation algorithm to adapt the NMF basis vectors for a speech signal. The proposed algorithm uses adaptation data to adapt the basis vectors twice. Bayesian adaptation is followed by a linear transformation of basis vectors. We applied the proposed adaptation algorithm to separate a speech signal from a background music signal when no enough training data for the speech signal that is in the mixture is available.

8. References

[1] Chin-Hui Lee and Qiang Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1241–1269, 2000.

[2] Ozerov A., Philippe P., Bimbot F., and Gribonval R., "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. of Audio, Speech, and Language Processing*, vol. 15, 2007.

[3] Tuomas Virtanen and A. Taylan Cemgil, "Mixtures of gamma priors for non-negative matrix factorization based speech separation," in *In Proc. of the 8th International Conference on Independent Component Analysis and Blind Signal Separation*, 2009.

[4] Tuomas Virtanen, "Spectral Covariance in prior distributions of non-negative matrix factorization based speech separation," in *EUSIPCO*, 2009.

[5] A.T. Cemgil, "Bayesian inference in non-negative matrix factorisation models," Tech. Rep., CUED/F-INFENG/TR.609, University of Cambridge, July 2008.

[6] Tuomas Virtanen, A. Taylan Cemgil, and Simon Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *In proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.

[7] C J Leggetter and P C Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–85, 1995.

[8] 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.

[9] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *in Proc. of ICASSP*, 2008.

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

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