Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 01, January-2019

Single Channel Speech Source Separation using Complex Matrix Factorization

P NagaYamini, Reg.1116468
Final Year M.Tech, Signal Processing
Sri Venkateswara University College of Engineering
Tirupati- 517502

Dr. T.Sreenivasulu Reddy
Dept. Of E.C.E,
Sri Venkateswara University College of Engineering
Tirupati - 517502

***Abstract*— Single Channel Speech Source Separation has been a challenging area of research in signal processing. Unlike many conventional methods of speech source separation that utilize information from multiple sensor readings, the problem of single source separation deals with separating sources obtained from only one sensor. Conventional methods of non-negative matrix factorization use only spectral magnitude and ignore spectral phase of the individual sources in the separation process. In this paper, we investigate the process of single channel speech source separation by considering a linear instantaneous model. A novel method of complex matrix factorization (CMF) that decomposes a complex spectrogram matrix into a complex base matrix and a real coefficient matrix is developed. Spectral phase of the sources are therefore incorporated in the process of source separation.**

The proposed separation method comprises of a learning followed by separation. In learning stage the basis vectors are obtained from the training data of the speakers taken from the corpus and a dictionary is created. In the separation stage, the dictionary is used to estimate the weights through complex matrix factorization in order to separate the sources. Experiments are performed using different mixture signals at various Signal-to-Signal Ratios (SSR) to evaluate the performance of the proposed method. The accuracy of source separation is measured by evaluating the objective measures like Log Likelihood Ratio (LLR) and Weighted Slope Spectral Distance (WSS). The proposed CMF method exhibits reasonably better performance when compared to other NMF methods and its variants available in literature.

I. INTRODUCTION

Source separation has been a topic of investigation for over two decades. The problem of source separation refers to the technique of separating the sources underlying in some mixtures of more than one source. A classical example of source separation is the cocktail party problem which represents the situation where a person is able to focus on a single conversation, when surrounded by a number of separate conversations. Separation can be classified as blind and non-blind. When no prior information about the sources is known, it comes under blind source separation. In contrast non-blind or supervised separation methods use prior information of the sources to train the separation model. Another type of classification of source separation methods is Over-determined source separation and Under-determined source separation, which is based on the number of sensors and sources. In over-determined case, the number of sensors is more than the number of sources and vice-versa in under-determined case.

Single channel source separation(SCSS) is the extreme case of under-determined source separation where only one mixture signal of more than one source signals, is available. In many practical applications, only one observation is available from the hardware and in such cases, conventional source separation techniques that require more than one observation are not appropriate. Hence the problem of single channel source separation has become widely interesting.

II. PROBLEM FORMULATION

Let z[n] be the mixed signal which comprises of two speakers $z_1[n]$ and $z_2[n]$. The problem of single channel speech source separation aims at obtaining the estimates of $z_1[n]$ and $z_2[n]$ using a single mixture signal z[n].

$$z[n] = z_1[n] + z_2[n]$$

We convert the time domain speech signal into frequency domain by finding the Short time fourier transform. Let

$$Z(k,\omega), Z_1(k,\omega), Z_2(k,\omega)$$

denote the STFT of z[n],$z_1[n]$ and $z_2[n]$ respectively where k represents the time frame index and represent the frequency bin index in the STFT domain. Therefore we can write

$$Z(k,\omega) = Z_1(k,\omega) + Z_2(k,\omega)$$

$$|Z(k,\omega)|e^{j\phi_{Z(k,\omega)}} = |Z_1(k,\omega)|e^{j\phi_{Z_1(k,\omega)}} + |Z_2(k,\omega)|e^{j\phi_{Z_2(k,\omega)}}$$

Various methods in literature use some training data and construct a set of basis vectors for all the sources present in the mixture. With the pre-learnt bases, weights are estimated for a mixture using matrix factorization methods from which sources can be separated. From each signal in the training set of clean speech, we extract a set of basis vectors X_{train}, which can be used in the separation process to calculate weights. Both learning the basis vectors and estimation of weights require complex matrix factorization. So the problem reduces to finding an accurate technique to estimate complex bases X_{train} and corresponding weights H_i such that

$$Z_i = X_{train} H_i$$

III. COMPLEX MATRIX FACTORIZATION APPROACH TO THE SPEECH SOURCE SEPARATION

Non Negative Matrix Factorization (NMF) is a linear basis decomposition technique, subject to constraints of non-negativity on data being imposed. It basically decomposes a non-negative matrix Z into product of two matrices X and H, constrained such that all the elements of X and H are non-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 01, January-2019

negative.The matrix X is termed as Basis vector matrix and H is termed as the Weight matrix or coefficient matrix, E is the residual matrix or simply the error in approximation.

$$Z = XH + E \approx XH$$

Cost functions measures the divergence between Z and XH. This can be expressed as below,

$$\{X, H\} = \arg\min_{X,H \geq 0} \{C(Z; X, H)\}$$

subject to $X, H >= 0$.

NMF is widely used for speech source separation. Decomposition of a mixed signal into corresponding basis vectors and estimation of corresponding weights is known to work well for single channel mixtures. In general NMF based separation assumes that the phase of the source signal is either equal to the mixed signal or it is assumed to be constant.

In the subsequent sections we propose a method of complex matrix factorization based source separation that includes the phase information also in the separation process thereby separation process is more accurate.

A. *Proposed Complex Matrix Factorization Method*

Consider Z to be a complex matrix which needs to be factorized into product of basis vector matrix X and weight matrix H,

Z = XH

where the base matrix X is complex and the weight matrix X is real.

The minimization problem is stated as

$$\{X, H\} = \arg\min_{X,H \geq 0} \{C(Z; X, H)\}$$

If the cost function used is Squared Euclidean Distance, then

$$\{X, H\} = \arg\min_{X,H} \left\{ \|Z - XH\|^2 \right\} \quad (3.0)$$

If the cost function used is KL-divergence, then

$$\{X, H\} = \arg\min_{X,H} \{D_{KL}(Z \| XH)\}$$

Let $\hat{Z} = XH$ be the approximated matrix. This factorization problem is a complex matrix factorization problem. Applying a simple transformation to convert complex matrix factorization problem into a non-negative matrix factorization problem and solve the task of source separation in NMF framework. The transformation is given as follows:

$$\hat{Z} = \hat{Z}_{+r} - \hat{Z}_{-r} + j(\hat{Z}_{+i} - \hat{Z}_{-i}) \quad (3.1)$$

where,

$$\hat{Z}_{+r} = \max(0, real(\hat{Z})), \quad \hat{Z}_{-r} = -\min(0, real(\hat{Z}))$$

$$\hat{Z}_{+i} = \max(0, imag(\hat{Z})), \quad \hat{Z}_{-i} = -\min(0, imag(\hat{Z}))$$

Here max, min, real and imag denotes element-wise functions operating on matrices. These functions calculates maxima, minima, real part and imaginary part of each element in the matrix. The complex matrices Z and X are separated by using the transformation defined in Eq.3.1. Hence,

$$Z = Z_{+r} - Z_{-r} + j(Z_{+i} - Z_{-i}) \quad (3.2)$$

$$X = X_{+r} - X_{-r} + j(X_{+i} - X_{-i}) \quad (3.3)$$

Weight matrix H is separated as

$$H = H_{+} - H_{-} \quad (3.4)$$

where

$$H_{+} = \max(0, real(H))$$
$$H_{-} = -\min(0, real(H))$$

max, real, min are the element-wise functions that operate on matrices and give the maximum, real part and minimum of the elements in the matrix.

Through the transformations defined above, we decompose a complex matrix into some non-negative matrices.

$$Z_{+r}, Z_{-r}, Z_{+i}, Z_{-i}, X_{+r}, X_{-r}, X_{+i}, X_{-i}, H_{+}, H_{-}$$

are the non-negative matrices that are obtained using the above transformations. Using equations 3.2, 3.3, 3.4 and substituting in $\hat{Z} = XH$ we get

$$\hat{Z}_{+r} = X_{+r}H_{+} + X_{-r}H_{-}$$
$$\hat{Z}_{-r} = X_{+r}H_{-} + X_{-r}H_{+}$$
$$\hat{Z}_{+i} = X_{+i}H_{+} + X_{-i}H_{-}$$
$$\hat{Z}_{-i} = X_{+i}H_{+} + X_{-i}H_{+}$$

Let,

$$\hat{Z}_1 = \hat{Z}_{+r}, \quad \hat{Z}_2 = \hat{Z}_{-r}, \quad \hat{Z}_3 = \hat{Z}_{+i} \quad \hat{Z}_4 = \hat{Z}_{-i}$$

Also for convenience let us have

$$Z_1 = Z_{+r} \quad Z_2 = Z_{-r} \quad Z_3 = Z_{+i} \quad Z_4 = Z_{-i}$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 01, January-2019

Now with all the notations described above, we convert complex matrix factorization into non-negative matrix factorization. On applying triangle inequality to equation 3.0, we obtain

$$\min_{X,H}\left\{\|Z-XH\|^2\right\} \le \min_{X,H}\sum_{k=1}^{4}\left\{\left\|Z_k-\hat{Z}_k\right\|^2\right\} \quad (3.5)$$

We know that Z_k, $\hat{Z}_k$ are independent of each other. Hence equation 3.5 becomes

$$\min_{X,H}\left\{\|Z-XH\|^2\right\} \le \sum_{k=1}^{4}\left\{\min_{X,H}\left\|Z_k-\hat{Z}_k\right\|^2\right\} \quad (3.6)$$

Hence the problem now reduces to

$$\min_{X,H}\left\|Z_k-\hat{Z}_k\right\|^2$$

for all

$$k \in \{1,2,3,4\}.$$

In fact the term on the R.H.S of the equation 3.6 represents the upper bound to the solution of the optimization problem in equation 3.0. Therefore convergence of R.H.S of the equation 3.6 guarantees convergence of the cost function in equation 3.0.

Sequential solving of these optimization problems will lead to a bias towards the first optimization problem. Hence we combine the sub-matrices into a single matrix and then solve them concurrently. This is shown as follows:

$$\begin{bmatrix}\hat{Z}_{+r} & \hat{Z}_{-r}\\ \hat{Z}_{+i} & \hat{Z}_{-i}\end{bmatrix} = \begin{bmatrix}X_{+r} & X_{-r}\\ X_{+i} & X_{-i}\end{bmatrix}\begin{bmatrix}H_{+} & H_{-}\\ H_{-} & H_{+}\end{bmatrix}$$

(or)

$$\begin{bmatrix}\hat{Z}_1 & \hat{Z}_2\\ \hat{Z}_3 & \hat{Z}_4\end{bmatrix} = \begin{bmatrix}X_1 & X_2\\ X_3 & X_4\end{bmatrix}\begin{bmatrix}H_1 & H_2\\ H_3 & H_4\end{bmatrix}$$

We denote the matrix on the L.H.S as Z_c and the matrices on the right side as X_c and H_c respectively.

Also, $X_1 = X_{+r}$, $X_2 = X_{-r}$, $X_3 = X_{+i}$, $X_4 = X_{-i}$.

Hence we have

$$\hat{Z}_c = X_c H_c$$

As $H_1 = H_4$ and $H_2 = H_3$, we perform an update after every NMF iteration as shown below.

$$H_1, H_4 \longleftarrow \frac{H_1+H_4}{2}$$

$$H_2, H_3 \longleftarrow \frac{H_2+H_3}{2}$$

Hence the CMF problem is reduced into NMF problem of the form

$$\min\left\|Z_c - X_c H_c\right\|^2$$

with respect to X_c and H_c,

where Z_c, X_c and H_c are non-negative matrices. Using the transformations, we have converted a complex matrix factorization problem into a non- negative matrix factorization problem. The approach described above is illustrated in figure 4.1.

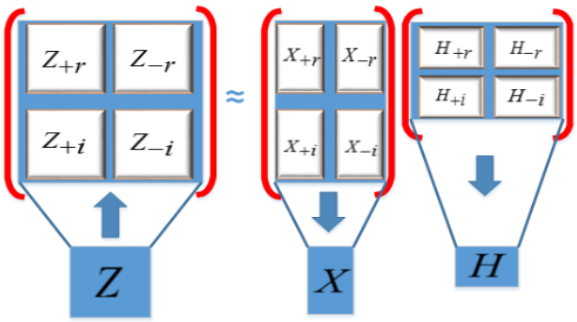

Fig. CMF Approach for joint modeling of Magnitude and Phase

A. Algorithm for computing X and H using CMF method

1: Input : Complex matrix Z and its transformed matrix Z_c.

2: Initialization : The matrices X_{+r}, X_{-r}, X_{+i}, X_{-i}, H_+ and H_- are assigned random non-negative values.

3: Rearrange the elements of these sub-matrices to form X_c and H_c as shown.

4: Alternating multiplicative updates : Update the elements of the sub matrices X_c and H_c by using the alternating multiplicative updates of NMF.

5: Update of weight matrix :

$$H_1, H_4 \longleftarrow \frac{H_1+H_4}{2} \quad \text{and} \quad H_2, H_3 \longleftarrow \frac{H_2+H_3}{2}.$$

6: Repeat : Step 4 to 5 for a number of iterations to minimize the error between Z_c and $\hat{Z}_c$.

7: Termination : Reconstruct the original matrices X and H from X_c and H_c by doing

$$X \leftarrow X_1 - X_2 + j(X_3 - X_4) \quad \text{and}$$
$$H \leftarrow H_+ - H_-$$

8: Output : Complex matrix X and real matrix H.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 01, January-2019

B. Dictionary Learning

In order to estimate $z_1[n]$ and $z_2[n]$, we first form over complete dictionaries that represent the basis vectors of speech of both the speakers. After extracting the basis vectors, we form the dictionary by concatenating the basis vectors of both the speakers. Then we decompose the mixed signal by using this dictionary and calculate corresponding weight matrix. The following algorithm describes the process of dictionary learning in case of two speakers speaking in one mixed signal.

C. Algorithm for Dictionary Learning

1: Input : Clean speech of the speakers, $z_1[n]$ and $z_2[n]$.

2: Short Time Fourier Transform (STFT) : Calculate Short Time Fourier Transform for each signal in the training data, say Z, which is a complex matrix.

3: Decomposition using transformations : Find the matrix Z_c from the complex STFT matrix Z using the transformations shown in figure 4.1.

4: Random initialization of basis vector matrix and weight matrix : The matrices

$$X_{+r}, X_{-r}, X_{+i}, X_{-i}, H_{+}, H_{-}$$

are assigned random non-negative values.

5: Rearrange the elements of these sub-matrices to form X_c and H_c.

6. Update the elements of the sub matrices X_c and H_c using the multiplicative update rules

7: Now find out the final non-negative matrices X_c and H_c .

8: Basis matrix extraction : The basis vector matrices X_c of all the speech signals of the same speaker are concatenated together to form the overall base matrix of that speaker.

9: Forming the dictionary : All the base matrices obtained in the previous step for all the speakers are concatenated together to form the overall dictionary X.

10: Output : Overall dictionary matrix X, which is used in the separation process.

D. Decomposition of the mixed signal

Using CMF the spectrogram (STFT) of the mixture signal Zmixture is decomposed into a product of complex matrix X and a real coefficient matrix H following the algorithm A but with a fixed basis matrix obtained from dictionary learning

$$Z_{mixture} = [(X_{basis})_1 (X_{basis})_2] H$$

The same equation after transformations turns out to be

$$Zc_{mixture} = [(Xc_{basis})_1 (Xc_{basis})_2] H_c$$

where $(Xc_{basis})_1$ and $(Xc_{basis})_2$ are the bases obtained from dictionary learning. It is fixed during the decomposition of the mixture spectrogram. Only coefficient matrix is updated using the update rule defined in algorithm A The basis matrix is kept fixed and H is initialised by random positive noise. The decomposition is done for a fixed number of iterations or until the convergence criterion is satisfied. After the decomposition, the spectrograms of the sources are estimated by converting X_c and H_c into X and H respectively by $X \leftarrow X_1 - X_2 + j(X_3 - X_4)$, $H \leftarrow H_+ - H_-$

E. Reconstruction of source signals

In the previous section, we have decomposed a complex STFT matrix into a product of complex base matrix and a real coefficient matrix. Now from the decomposed matrices we need to estimate the complex STFT of individual sources. The weight matrix H can be split as two sub-matrices H1 and H2, each one containing the coefficients for a particular speaker in the mixture. Hence, the estimated complex STFT of the source signals in the mixture are given as,

$$Z_1 = (X_{basis})_1 H_1$$

$$Z_2 = (X_{basis})_2 H_2$$

where $(X_{basis})_1$ and $(X_{basis})_2$ are the base matrices of the speaker 1 and speaker 2, extracted using Algorithm C, are the complex STFT of the sources present in the mixture. Now to reconstruct the signals in time domain, take the Inverse short time fourier transform (ISTFT) of and estimate the time domain source signals.

IV. PERFORMANCE EVALUATION

A. Database

For all the experiments, the audio signals taken from GRID corpus[6] which is an audio-visual corpus used in speech perception and automatic speech recognition studies. The corpus consists of high-quality audio and video recordings of 1000 sentences spoken by each of 34 talkers (18 male and 16 female). Each speech signal consists of sentences of the form ”bin blue at L 6 please”. It is actually ”< command :1>< color :1>< preposition : 1 >< letter : L >< digit:6>< adverb :3>”.

The data set consists of single channel speech signals in ”.wav” format with a sampling frequency of 25 kHz. Corpus actually consists of video files of the speakers but our interest is only on speech source separation. Hence we have taken only the audio files. The complete corpus and transcriptions are freely available for research use.

B. Evaluation Criteria

To do the performance analysis and to measure the performance of the proposed method for single channel speech source separation, both subjective and objective quality measures exist in literature. Subjective measures are obtained by collecting the rating by group of listeners. Evaluation is done by using human auditory system and its perception to re-constructed audio. The main disadvantage of these subjective measures is that they are time-consuming and expensive but they have high validity. The second way of validation criteria is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 01, January-2019

by evaluating objective measures which extract a metric of speech quality between the reconstructed speech signal and the original reference signal using mathematical techniques. In our experiments we use objective measures in our performance analysis.

C. Objective Evaluation Methods

In objective evaluation measures, we assess the speech quality from the extracted physical parameters of the reconstructed speech signals. The reference speech signals which are used to form the mixed signals are taken and they are compared with the reconstructed speech signals after source separation. Objective measures are used to measure the improvement of speech quality before and after separation of source signals. In this section we describe the objective measures that are used to compare the quality of source separation. We used Linear Predictive Coefficients based measures - Log Likelihood Ratio (LLR) and Weighted Slope Spectral Distance (WSS) for performance analysis of our separation process. Lower the values of LLR and WSS better is the performance.

Experiments are performed using the audio files taken from the GRID database mentioned. Audio files of two speakers, one male and one female are taken from the corpus. The dataset is divided into two parts, training dataset and testing dataset. Used 300 speech audio files of both the speakers for training the bases. 10 mixture signals from the same speakers of Signal-to-Signal ratios (SSR) 0dB,4dB,8dB and 10dB are taken for the testing the accuracy of separation. From each signal in the training data set, 10 basis vectors are extracted. This number is variable and set by the user. Separation is done by Non-negative matrix factorization method and Complex matrix factorization method Results obtained through various subjective and objective measures are presented in the following sections. As mentioned the results are compared with existing techniques of source separation using Non-negative matrix factorization and tabulated properly. In all the experiments, fixed the window size to be 800 samples(32ms at 25khz), 50% overlap between adjacent frames and we take 512 point DFT for each frame. Experiments are carried out for different values of from 0.1 to 1 in steps of 0.1 and for = 2, 5 and 10. Used randomly generated mixture signals of different signal to signal ratios (SSR) from the grid corpus. Experiments are also conducted where is varied for every iteration.

D. Results Obtained using Log Likelihood Ratio(LLR)

Lower the value of LLR, better is the performance. As observed in table 5.2, the LLR values are lower for our CMF than NMF.

Methods	Speaker	SSR = 0dB	SSR= 4dB	SSR= 8dB	SSR= 10dB
NMF	speaker1	1.13	1.84	1.19	1.86
	speaker2	1.10	1.42	1.23	1.45
CMF	speaker1	0.80	0.75	0.80	0.85
	speaker2	0.79	0.77	0.79	0.81

Table 5.2: Objective measures results using LLR at different SSRs

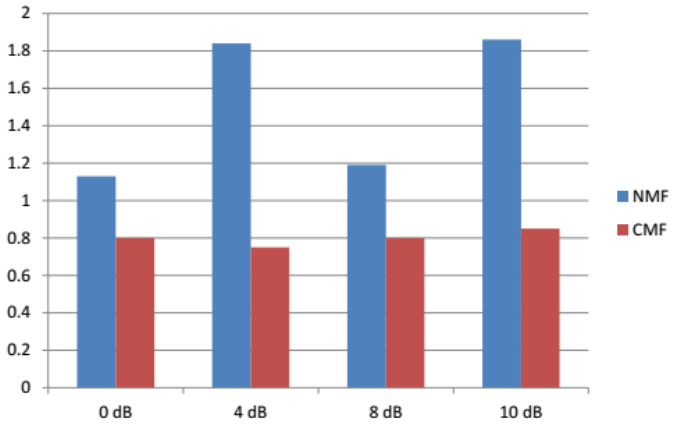

Figure 5.1: LLR of speaker 1 at different Signal-to-signal ratios (SSR)

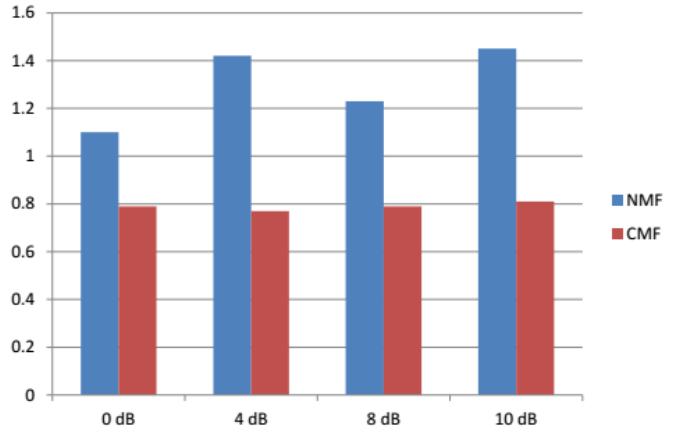

Figure 5.2: LLR of speaker 2 at different Signal-to-signal ratios (SSR)

E. Results Obtained using Weighted Slope Spectral Distance (WSS)

Lower the value of WSS, better is the performance. As observed in table 5.3, the WSS values are lower for our CMF than NMF.

Methods	Speaker	SSR = 0dB	SSR= 4dB	SSR= 8dB	SSR= 10dB
NMF	speaker1	10.53	9.84	9.56	9.43
	speaker2	11.10	11.71	11.88	12.28
CMF	speaker1	6.55	6.81	6.42	6.91
	speaker2	10.02	10.77	10.33	10.59

Table 5.3: Objective measures results using WSS at different SSRs

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 01, January-2019

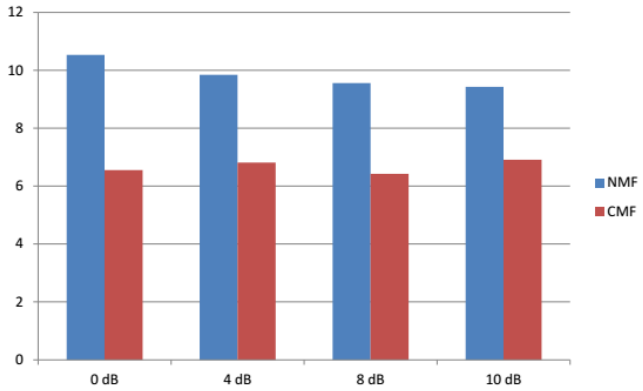

Figure 5.3: WSS of speaker 1 at different Signal-to-signal ratios (SSR)

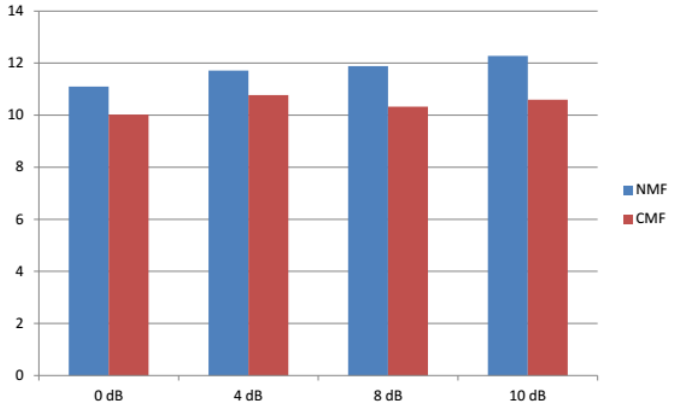

Figure 5.3: WSS of speaker 2 at different Signal-to-signal ratios (SSR)

V. CONCLUSION

This paper proposed a novel method for single channel speech source separation using complex matrix factorization. In CMF, decomposed a complex matrix factorization problem into a non-negative matrix factorization problem and hence able to model both magnitude and phase of the short time fourier transform (STFT) of the speech waveforms. Hence this method overcomes the disadvantage of the traditional source separation techniques by Non-negative matrix factorization that utilizes only magnitude information neglecting the phase information. The detailed procedure for the conversion of a CMF problem into NMF problem is discussed in the chapters covered previously. All the experiments are done using the GRID corpus. The performance analysis is also done and proposed technique is compared with the existing techniques that operate only on magnitude ignoring the phase information. Significant improvement is observed for the proposed method when compared to existing methods. This work can be further extended by considering the noisy mixture signal recorded in reverberant surroundings. Therefore this can be carried out in real time scenarios where our proposed method can be implemented and is expected to give better results for single channel source separation.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in In NIPS, pp. 556–562, MIT Press, 2000.

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

[3] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization with sliding windows and spectral masks.," in INTERSPEECH, pp. 1773–1776, ISCA, 2011.

[4] J. Eggert and E. Korner, "Sparse coding and nmf," in Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, vol. 4, pp. 2529–2533 vol.4, 2004.

[5] Z. L. Zunyi Tang, Shuxue Ding and L. Jiang, "Dictionary learning based on nonnegative matrix factorization using parallel coordinate descent," Abstract and Applied Analysis, vol. 2013, 2013.

[6] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," The Journal of the Acoustical Society of America, vol. 120, pp. 2421–2424, November 2006.

[7] W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on, vol. 2,pp. 749–752 vol.2, 2001.

[8] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 1, pp. 229–238, 2008.

[9] R. Crochiere, J. Tribolet, and L. Rabiner, "An interpretation of the log likelihood ratio as a measure of waveform coder performance," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 3, pp. 318–323, 1980.