A VQ-Based Single-Channel Audio Separation for Music/Speech Mixtures.

CONFERENCE PAPER · JANUARY 2009

DOI: 10.1109/UKSIM.2009.123 · Source: DBLP

4 AUTHORS, INCLUDING:

Meysam Asgari
Oregon Health and Science University
16 PUBLICATIONS **35** CITATIONS
SEE PROFILE

Elahe Aboueimehrizi
Imperial College London
1 PUBLICATION **2** CITATIONS
SEE PROFILE

Ali Mostafavi
4 PUBLICATIONS **3** CITATIONS
SEE PROFILE

A VQ-based Single–Channel Audio Separation for Music/Speech Mixtures

Meysam Asgari[1], Mahdi Fallah[2]
Dept. of Electrical Engineering, Amirkabir University of Technology (Tehran Polytechnic)
Tehran, Iran
e-mail: [1]meysam_asgari@aut.ac.ir,
[2]fallah_moafi@aut.ac.ir

Elahe Abouie Mehrizi[3], Ali Mostafavi[4]
[3]Dept. of Mechanical Engimeering, [4]Dept. of Electrical Engineering, University of Birmingham
Birmingham, UK
e-mail: [3]elahe8607@yahoo.com,
[4]alimostafavi@rocketmail.com

Abstract— *In this paper, we address the problem of audio source separation with one single sensor, based on estimation of statistical model of the sources. We improve the-state-of-the art Vector Quantization (VQ) by considering apriori histograms of huge training data. This will result in a more accurate codebook for each source in contrast to the commonly used Linde-Buzo-Gray (LBG) algorithm. An optimum estimator is introduced in separation stage based on Discrete Fourier Transform (DFT) amplitudes. Finally, conducting different simulations, it is demonstrated that proposed approach efficiently segregated audio mixtures in terms of Signal to Distortion Ratio (SDR) measures as well as Mean Opinion Score (MOS) criterion.*

Keywords- Single Cahnnel Audio Sepatation; Vector Quantization

I. INTRODUCTION

Single channel sound separation has been introduced as a challenging topic in past decade. Although new tools, such as independent component analysis (ICA) have been proposed, developed, and improved like in [2],[5],[7],[12] and good results are reported by these methods, they are restricted whenever number of microphones is larger or equal to the number of sources and they significantly fail when employed in the case of monaural sound separation [11].

Many algorithms have been proposed for single channel audio separation and as a whole they can be divided into two distinct categories: Computational Auditory Scene Analysis (CASA) and statistical spectral models. CASA aims at identifying source signals by grouping some related psychoacoustic cues (e.g. notes in music and onset, offset, and harmonicas for speech) and grouping them into streams [1]. CASA has some detrimental drawbacks including: the precedence rules between grouping cues are sometimes hard to assess and the correlogram front-end used by most methods prevents identification of masked auditory objects and separation of harmonic partials cited within same critical band. Moreover existing methods for multichannel recordings are restricted to non-reverberant mixtures. In addition they can hardly segregate instruments playing in the same pitch range into different streams and usually result in crosstalk in the separated output signal.

As the second group, statistical spectral models have been proposed for the separation of mono recordings. In [3], each target source is modeled as the sum of elementary components with known power spectral densities (PSDs). The approach involves a non negative decomposition of the spectra of the observed mixture in a given frame into a dictionary of PSDs. The resolution of the PSDs varies at each iteration of the algorithm. However, as reported in [3], splitting the signal in source components and a residual component is a rather difficult task.

As another method, Independent Subspace Analysis (ISA) decomposes the mixture power spectrogram as a sum of typical spectra with time-varying weights, builds the source power spectrograms by grouping these spectra into subspaces and computes the source waveforms by inverting their spectrograms [13] or by adaptive Wiener Filtering [15]. Typical spectra are estimated from the mixture using Non negative Matrix Factorization (NMF) [16]. Good results were reported for note transcription on solo recordings [17], [16], however ICA and NMF badly separate low-intensity notes [18] and produce spurious notes with short duration, and their ability to segregate non-percussive instruments has not been studied. Hidden Markov Models (HMM) solve these issues by learning accurate priors for the log-power spectra of the sources on solo data and by setting a prior on event duration. Satisfying separation results were obtained on speech mixtures with factorial combination of source models [19]. But complex parameter sharing procedures are needed on musical mixtures to avoid over learning [20], since the number of hidden states for each source (i.e. the number of chords it can play) may be very large.

In this paper, by incorporating the so called Vector Quantization (VQ) algorithm with an optimum estimator, a novel monaural audio separation method is proposed. The paper is organized as follows. In the following section, we present a brief review on existing approaches used for single-channel audio separation. In section 3, an optimum estimator is proposed. Section 4, presents some simulation results and section 5 concludes.

II. EXISTING METHODS

Consider a mixture consisting of speech and music signals as:

$$\boldsymbol{z}(n) = \boldsymbol{m}(n) + \boldsymbol{s}(n) \qquad n = 1,...,N \tag{1}$$

where ***m(n)*** is the music signal, ***s(n)*** denoted the speech and ***z(n)*** the resulting mixture. Our proposed approach is to establish some apriori stochastic model for the underlying sources. This modeling can be accomplished by several approaches including *Gaussian Mixture Model* (GMM) [1], *Gaussian Scaled Mixture Model* (GSMM) [10-11], HMM [6],[8],[19],[20] and VQ [4],[9]. Among these approaches, GMM has been considered as favorable choice in music speech application. By employing *short-Time Fourier Transform* (STFT) the mixture frame can be written as:

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

where n denoted the frame number while f is the frequency index in a time-frequency representation. Assuming a diagonal covariance matrix $\sum_m = diag(\sigma_m^2(f))$, $\sum_S = diag(\sigma_s^2(f))$ for speech and music, respectively, in mixture mode in (1) Bayesian estimation will be [11]:

$$\hat{\boldsymbol{M}}(n,f) = \frac{\sigma_m^2(f)}{\sigma_m^2(f)+\sigma_s^2(f)}\boldsymbol{X}(n,f) \tag{3}$$

$$\hat{\boldsymbol{S}}(n,f) = \frac{\sigma_S^2(f)}{\sigma_m^2(f)+\sigma_s^2(f)}\boldsymbol{X}(n,f) \tag{4}$$

It is seen that the more mixtures are used, the more accurate the resulting model will be as reported in [21]. Such GMM model may introduce some over-estimation error as well as introducing some interference signal degrading separated perceptual quality while used to model the sources. This usually occurs, when some interference signal remain in the separated output signal i.e. low *Signal to Interference Ratio* (SIR). Cope these problems, in this paper, a modified version of VQ modeling is used to establish source models. In addition, the GMM method is limited to employ mixture phase which results in bad separated output signals. However, in the proposed VQ based approach, we insert some phase vectors at each VQ entries which are to be used only in the synthesis stage to produce separated signals with natural quality.

III. AUDIO SOURCE STOCHASTIC MODLING

A. Feature Selection

In order to have a good stochastic model, one need to gather a huge data samples from audio source. In addition, the feature type plays an important role in source modeling. The more compact a feature density (lower variance), a better audio source model may result. STFT is commonly used as appropriate feature in previous works. However, it contains both amplitude and phase information. Due to uniformity of phase distribution [21, 22], this STFT is not capable for clustering purposes in its raw format. As a result, by neglecting the phase information, we choose spectrum amplitude as our feature vectors.

B. Vector Quantization

Constructing related codebooks for each audio source requires some feature vectors which are obtained by employing STFT analysis. For clustering the so-called LBG algorithm is used. As reported in [9], the initialization stage plays a key role in clustering overall performance in that it does not guarantee to reach at a global minimum. As a result, to cope this problem, some histogram sensitive approach is used in this paper. This histogram information can be established as follows. Starting from the 1st input vector, we scan for those training vectors having Euclidean distance less than a pre-defined threshold (also called discrepancy threshold which is defined by practical evaluations). Finding all vectors closest to the 1st training vector, we insert the number of these vectors in a temporary vector namely histogram index vector. Assuming N be the number of training vectors, by repeating the procedure for other N vectors, we complete histogram vector entries. Next, this vector is sorted decreasingly and finally, first M number of indices is selected. The related vectors to these indices are mostly probable in terms of histogram information.

In addition, only frequency range within [0,1 kHz] is perceptually important in the obtained spectrum amplitude vectors for audio signals. As a result, some weighting distance can be useful in the proposed VQ approach. This weight attenuates frequencies over 1 kHz as depicted in Fig.1.

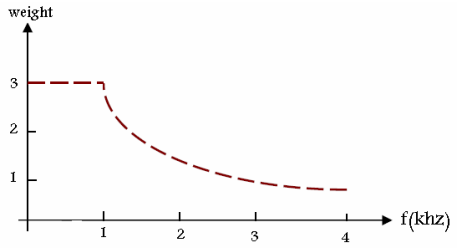

Figure 1. Weight function to improve DFT amplitude clustering capability.

Employing such weighting as well as histogram initialization procedure explained earlier, we result in a efficient audio model for speech and music sources.

IV. SEPARATION ALGORITHM

DFT amplitude was selected as our feature vector as explained in the previous section. The STFT amplitude of the mixture signal is equal to the underlying source STFTs'. By taking STFT from (1), we have:

$$\boldsymbol{X}(f) = \boldsymbol{M}(f) + \boldsymbol{S}(f) \tag{5}$$

where *M(f), S(f)* and *X(f)* are the STFT representation for the underlying speech, music and mixture, respectively and are expressed as follows:

$$\begin{aligned}\boldsymbol{X}(f) &= [x_1 e^{j\theta x_1}, ..., x_k e^{j\theta x_k}, ...x_N e^{j\theta x_N}] \\ \boldsymbol{M}(f) &= [m_1 e^{j\theta m_1}, ..., m_k e^{j\theta m_k}, ...m_N e^{j\theta m_N}] \\ \boldsymbol{S}(f) &= [s_1 e^{j\theta s_1}, ..., s_k e^{j\theta s_k}, ...s_N e^{j\theta s_N}]\end{aligned} \quad (6)$$

and the related spectrum amplitudes are:

$$\begin{aligned}|\boldsymbol{X}(f)| &= [x_1, ..., x_k ., ...x_N] \\ |\boldsymbol{M}(f)| &= [m_1, ..., m_k, ...m_N] \\ |\boldsymbol{S}(f)| &= [s_1, ..., s_k, ...s_N]\end{aligned} \quad (7)$$

while the observed mixture signal depend on both amplitude and phase of the underlying DFT spectra. In order to find a mixture estimator we can form the probability distribution of amplitude for mixture as follows:

$$\begin{aligned}|\boldsymbol{X}(f)| = [&|m_1 e^{j\theta m_1} + s_1 e^{j\theta s_1}|, .. \\ &.., |m_k e^{j\theta m_k} + s_k e^{j\theta s_k}|, ..., |s_N e^{j\theta s_N} + m_N e^{j\theta m_N}|]\end{aligned} \quad (8)$$

It is proven that for audio signals, especially for speech signals, the phase distribution is well approximated by a uniform distribution [21,22]. Now considering that phase related distribution for speech and music be uniform, denoted as $\theta m_k, \theta m_k$, the mixture DFT amplitude can be written as follows:

$$x_k = |m_k e^{j\theta m_k} + s_k e^{j\theta s_k}| \quad (9)$$

using Maximum Likelihood estimator for estimation xk we get:

$$\hat{x}_k = \max p_{x_k}(x_k) \quad (10)$$

due to the randomness of both θm_k and θs_k, the mixture signal, x_k is also a random variable. The *probability density function* (pdf) of the mixture can be obtained by using the training data. As a result, the amplitude histogram x_k for is illustrated in Fig.2 as an estimation for $P_x(x_k)$.

As it is observed in Fig.2,3, m_k+s_k has the most frequency of occurrence and the following estimation will be the best which is confirmed by Maximum Likelihood estimator given in (10) as follows:

$$\hat{x}_k = \max p_{x_k}(x_k) = m_k + s_k \quad (11)$$

and for the mixture signal we have:

$$|\boldsymbol{X}(f)| = |\boldsymbol{M}(f) + \boldsymbol{S}(f)| \approx |\boldsymbol{M}(f)| + |\boldsymbol{S}(f)| \quad (12)$$

As a result, the optimum approximation for DFT amplitude is equal to the addition of each of the underlying spectrum amplitudes. This estimator is called Maximum Likelihood Amplitude Estimator (MLAE). To establish audio models i.e. .|S(f)| and |M(f)| as speech and music code-vectors are established. Next, we search in each of the codebooks of the related sources for those indices which result in the closest approximation of the mixture signal. This can be best shown in the following equation:

$$i, j = \underset{i^*, j^*}{\arg\min} \left\{ |\boldsymbol{X}(f)| - \left(|\boldsymbol{M}_{i^*}(f)| + |\boldsymbol{S}_{j^*}(f)| \right) \right\} \quad (13)$$

where *i,j* are the codebook indices which are obtained from (6). Using the related codevectors to *i,j* the mixture vectors can be separated. Note that these codevectors only consist of DFT amplitude information and not the phase. Other approaches including PSD and *Mixture Maximization* (MIXMAX) has been used in [15],[22], respectively as follows:

$$i, j = \underset{i^*, j^*}{\arg\min} \left\{ |\boldsymbol{X}(f)|^2 - \left(|\boldsymbol{M}_{i^*}(f)|^2 + |\boldsymbol{S}_{j^*}(f)|^2 \right) \right\} \quad (14)$$

$$\begin{aligned}i, j = \underset{i^*, j^*}{\arg\min} \{ &\log(|\boldsymbol{X}(f)|) - ... \\ &.. \max\left(\log(|\boldsymbol{M}_{i^*}(f)|), \log(|\boldsymbol{S}_{j^*}(f)|) \right) \}\end{aligned} \quad (15)$$

As you can see, MIXMAX only considers the maximum of amplitudes which can not be the optimum estimation of x_k as illustrated in Fig.1,2. In fact, MIXMAX or PSD introduce approximately them median value in the range between ($|s_k-m_k|, |s_k+m_k|$)|which is not necessarily optimum estimation compared with proposed estimator in (10). For comparison purposes, the proposed Maximum Likelihood Amplitude Estimator (MLAE) approach is evaluated with these methods. Fig.4. demonstrate the whole separation procedure used in this work.

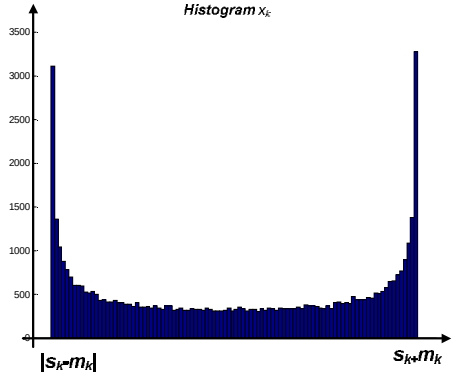

Figure 2. Amplitude histogram for mixture signal for distant related amplitude signals s_k and m_k.

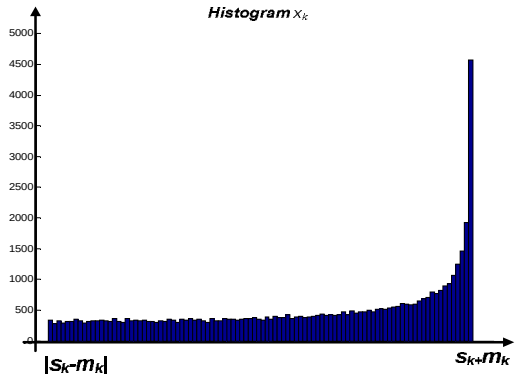

Figure 3. Amplitude histogram for mixture signal for closely related amplitude signals s_k and m_k.

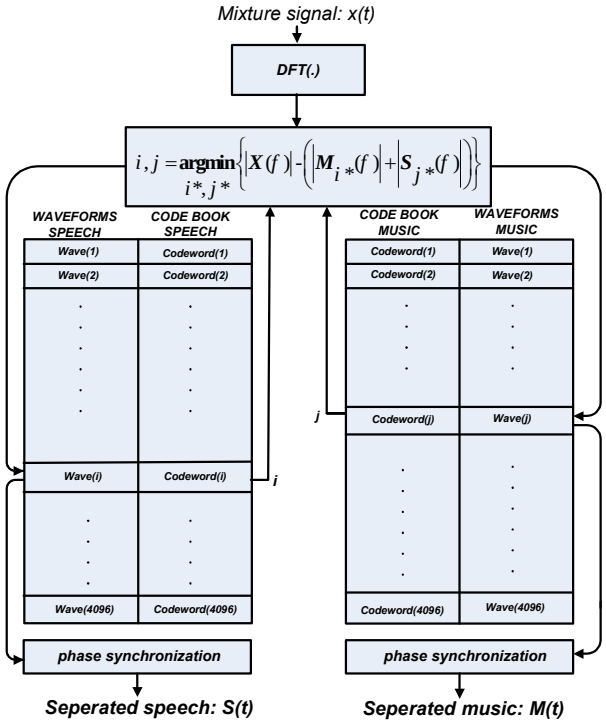

Figure 4. Whole segregation procedure.

V. SIMULATION RESULTS

For evaluating different algorithms with our proposed method, we have used 100,000 training vectors for 80 sentences uttered by 2 male/female speakers as well as music signals including piano segments extracted from Motzart audio waves. The speaker utterances are gathered from 230 different sentences which are phonetically balanced. The analysis window is set to 32 ms and a frame shift of 16 ms is used for a sampling frequency f_s=8 kHz. The codebook size is 12 bit or M=4096 clusters. To evaluate the separation performance, *Signal to Distortion Ratio* (SDR) value is considered as a measure to ensure how much interfering signal is separated from the mixture. Note that it is not necessary to compute SIR values which is reported in some previous works since the proposed method does not use the mixture signal in the synthesis stage but employs the resulting codebook indices instead. This SDR criterion is calculated in the DFT amplitude domain and not the time domain due to its dependency to phase. As a result, the SDR criterion can be defined as follows:

$$\mathrm{SDR}_i = 10\log_{10}\frac{\sum_{f=1}^{L}\boldsymbol{X}_i(f)^2}{\sum_{f=1}^{L}\left(\boldsymbol{X}_i(f)-\hat{\boldsymbol{X}}_i(f)\right)^2} \qquad (16)$$

where $\hat{\boldsymbol{X}}_i(f)$ and $\boldsymbol{X}_i(f)$ are the original and the reconstructed signal DFT amplitudes, respectively. Taking average over all frames we have:

$$\mathrm{ASDR} = \text{average}\,(\mathrm{SDR}_i) \qquad (17)$$

Table 1 summarizes the obtained *Average SDR* (ASDR) results for music/speech separation in the case of using different methods including MLAE, MIXMAX and PSD. As it is seen, ASDR of the proposed MLAE outperforms from other methods about 1 dB with respect to MIXMAX and 2 dB from PSD approach. Fig.5 depicts the mixture speech/music as well as separated underlying signals both in time and spectrogram domains. As our subjective results, we conducted Mean Opinion Score (MOS) to evaluate the output separated signals in terms of perception. As it seen, our proposed method outperforms from other previous ones including PSD and MIXMAX about 2 to 1 dB in terms of MOS

VI. CONCLUSION

In this paper an improved version of the so-called VQ clustering algorithm was proposed for audio single channel separation scenario. To segregate mixed audio frames a novel optimum estimator namely MLAE was employed. It was demonstrated that the proposed approach outperformed compared with other previous MIXMAX or PSD estimators in terms of SDR results as well as MOS results. As a result, the proposed approach was successfully removes SIR problem existent in common audio separation techniques introduced so far.

TABLE I. ASDR FOR SEPARATED SPEECH AND MUSIC SIGNALS USING MLAE, MIXMAX AND PSD ALGORITHMS.

Signal	!!PSD	MIXMAX	MLAE
Speech	*4.1772*	*5.0788*	*5.9965*
Music	*5.9133*	*7.125*	*7.9133*

TABLE II. MOS RESULTS FOR SYNTHESIZED OUTPUT SIGNALS.

Method	Category (Speech/Music)	MOS
MLAE	*Adult male + music*	*3.3*
	Adult female + music	*3.2*
MIXMAX	*Adult male + music*	*2.5*
	Adult female + music	*2.4*
PSD	*Adult male + music*	*2.2*
	Adult female + music	*2.1*

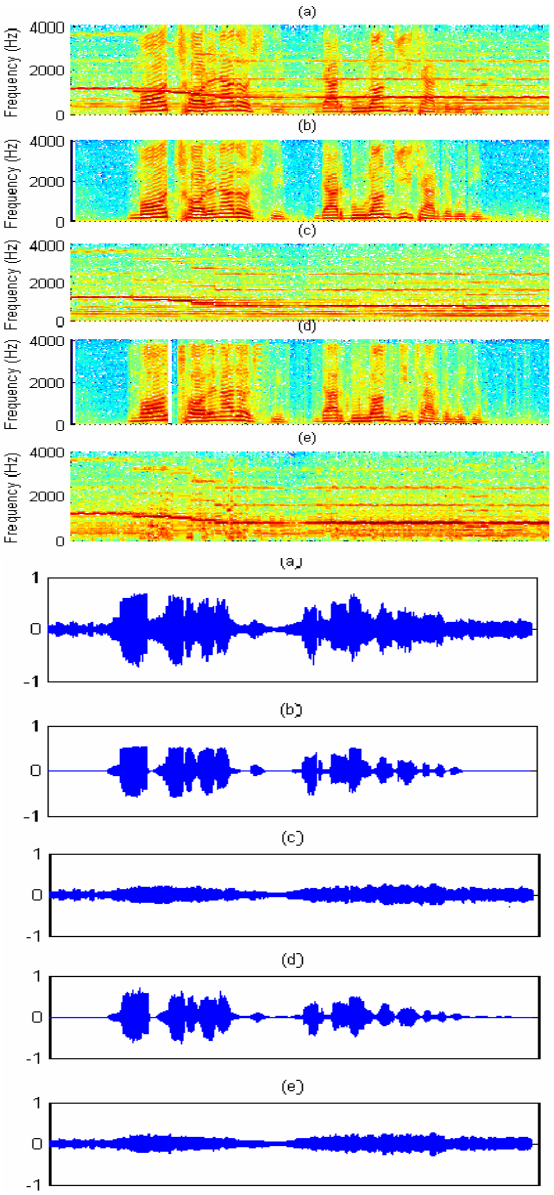

Figure 5. Corresponding spectrograms and time domain of (a) speech/music mixture, (b) underlying speech, (c) music, (d) separated speech, (e) separated music signal.

REFERENCES

[1] D. Ellis, "Prediction-driven computational auditory scene analysis", *Ph.D. dissertation*, MIT, 1996.

[2] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.

[3] L. Benaroya, R. Blouet, C. Févotte, and I. Cohen., "Single sensor source separation using multiple-window STFT representation, " *In Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC'06)*, Paris, France, Sep. 2006.

[4] D. Ellis and R. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation", in *Proc. ICASSP-06*, May 2006, vol. V, pp. 957–960.

[5] O. Bermond and J.-F. Cardoso, "Approximate likelihood for noisy mixtures", in *Proc. ICA'99*, Aussois, France, 1999, pp. 325–330.

[6] M. Reyes-Gomez, B. Raj, and D. Ellis, "Multi-channel source separation by factorial HMMs", in *Proc. ICASSP*, 2003.

[7] J. F. Cardoso, "Blind signal separation: Statistical principles" , *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[8] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMMs to segment models: a uni_ed view of stochastic modeling for speech recognition", *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 5, 1996.

[9] A. Gersho and R. M. Gray, "Vector Quantization and Signal Compression", *Kluwer Academic*, Norwell MA, 1992.

[10] C. Fevotte , S.J. Godsill, "A Bayesian approach for blind separation of sparse sources", *IEEE Trans. Speech Audio Process*, Vol. 4 (99), pp. 1–15, 2005.

[11] Laurent Benaroya, Frédéric Bimbot, and Rémi Gribonval , "Audio Source Separation With a Single Sensor" *IEEE Transaction on Audio, Speech, and Language Processing*, VOL. 14, NO. 1, JANUARY 2006

[12] T. Lee, M. Lewicki, M. Girolami, and T. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Processing Lett.*, vol. 6, no. 4, pp. 87–90, Apr. 1999.

[13] M. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis",. in *Proc. ICMC*, 2000.

[14] P. Mowlaee, A. Sayadian, "A Fixed Dimension Modified Sinusoid Model (FD-MSM) for Single Microphone Sound Separation", *in proceeding ICSPC*, Dubai, United Arab Emirates, 24-27 November, 2007.

[15] M. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis", in *Proc. ICMC*, 2000.

[16] L. Benaroya, L. McDonagh, F. Bimbot, and R. Gribonval, "Non-negative sparse representation for Wiener based source separation with a single sensor", in *Proc. ICASSP*, 2003.

[17] Kristjansson, T., Attias, H., Hershey, J., 2004. "Single microphone source separation using high resolution signal reconstruction", in *Proc. IEEE Int. Conf. Acous., Speech, and Sig. Proc., Montreal, ICASSP*-04, pp. 817–820, May, 2004.

[18] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription", in *Proc. WASPAA*, 2003.

[19] S. Abdallah and M. Plumbley, An ICA approach to automatic music transcription", in *Proc. 114th AES Convention*, 2003.

[20] S. Roweis, "One microphone source separation", in *Proc. NIPS*, 2000.

[21] H. Pobloth and W. B. Kleijn, "Squared error as a measure of perceived phase distortion", *J. Acoust. Soc. Am.*, vol.114, no. 2, pp. 1081–1094, Aug. 2003.

[22] D.Burshtein and S.Gannot,"Speech enhancement using a mixture-maximum model", *IEEE Trans. Speech and Audio Pr--ocessing*, vol. 10, no. 6, pp. 341–351, Sept. 2002.