IJICT International Journal of Information & Communication Technology
ITRC
Volume 2- Number 1- May 2010 (pp.1-10)

VQ-based Approach to Single-Channel Audio Separation for Music and Speech Mixtures

Pejman Mowlaee
Electrical Engineering Department
Amirkabir University of Technology
Tehran, Iran
pmowlaee@aut.ac.ir

Abolghasem Sayadiyan
Electrical Engineering Department
Amirkabir University of Technology
Tehran, Iran
eea35@aut.ac.ir

Hamid Sheikhzadeh Nadjar
Electrical Engineering Department
Amirkabir University of Technology
Tehran, Iran
hsheikh@aut.ac.ir

Received: February 8, 2009- Accepted: April 19, 2010

Abstract— **In this paper, we propose a low-complexity model-based single-channel audio separation approach. The proposed method presents three certain advantages over previous methods: 1) replacing commonly used linear masks like Wiener filtering by a proposed non-linear one, we show that it is possible to lower the crosstalk of the interfering source often occurring in a mask-based method while recovering the underlying signals from the observed mixture. Using nonlinear masks establishes a tradeoff between acceptable level of interference and low speech distortion, 2) as a post-processing stage, we use phase synchronization technique to enhance the perceptual quality of the re-synthesized signals, and 3) the proposed method is based on vector quantization (VQ) codebooks. Hence, the complexity is lower than previous GMM-based methods. Through extensive experiments, it is demonstrated that the proposed method can achieve a lower signal-to-distortion ratio (SDR). According to our listening experiments and according to the Mean Opinion Score (MOS) results, it is confirmed that the proposed method is able to recover separated outputs with a higher perceived signal quality.**

Keywords- vector quantization; nonlinear mask;audio source separation;model-based method; signal-to-distortion ratio.

I. Introduction

Single channel audio source separation has been introduced as a challenging topic in recent years. The audio source separation scenario is categorized into two main groups: 1) multi-channel, and 2) single-channel scenario. In multi-channel scenario, independent component analysis (ICA) has been widely used [1-4] and good results have already been reported. The separation performance of an ICA-based method, however, is confined and restricted when the number of microphones is larger or equal to the number of sources. In single-channel scenario, the goal is to estimate the separated signals according to their mixture recorded by a single microphone. Mathematically, this is equivalent to solve a linear equation with one observation (observed mixed signal) and two unknowns (the unknown underlying signals forming the observed mixture). The problem, in general, is not solvable without any *a priori* knowledge of the underlying signals. As one important auxiliary information, model-based separation methods employ statistical models to model the underlying signals and use this model for separating the signals from their mixture. In this paper we only focus on model-based single-channel scenario and the signals of interest are audio signals.

As the most representative approach for single-channel audio separation, model-based methods have widely been used. For instance, in [5], each one of the underlying desired sources are first being 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 used in separation varies in different iterations to another during the algorithm. According to [3], splitting the signal into its source components and a residual component is a rather difficult task. In [6], independent subspace analysis (ISA) was employed to decompose the power spectrogram of a given audio mixture into a sum of spectra with time-varying weights. Then the source power spectrograms are reconstructed by grouping the weights into subspaces and computing the source waveforms by employing a Wiener filtering framework. In [7], an ICA-based approach was employed for single-channel audio separation. It was observed that the method results in an inferior performance for speech + music mixtures. More specifically, it was observed in [7] that the ICA-based method could recover mixture of two music signals more cleanly than a mixture of speech and music signal. In [8],[9], non-negative matrix factorization (NMF) was introduced as a way to estimate the spectra of the unknown sources according to the observed mixed signal. Rather good separation results were reported for note transcription on solo recordings [8],[10]. In [9], the performance of the ICA and NMF were studied for audio source separation. According to [11], both ICA and NMF are not capable to appropriately separate the low-intensity notes. These methods produce spurious notes with short duration and their ability to segregate non-percussive instruments has not been studied. In [12], this issue was considered by employing Factorial hidden Markov models (HMM). By applying HMM, accurate priors for the log-power spectra of the sources could be learnt on solo data. According to [13], satisfying separation results were obtained on speech mixtures with factorial combination of source models. However, complex parameter sharing procedures are required on musical mixtures to avoid overlearning, since the number of hidden states for each source (equivalently the number of chords it could play) may be very large.

The statistical methods already used for the model-based audio separation are categorized into HMM [12-15], Gaussian mixture model (GMM), recently Gaussian scaled mixture model (GSMM) [16],[17], and VQ [18],[19]. Among these approaches, the GMM-based approach has been of more interest and has been introduced as an attractive candidate for audio source separation. According to [20], the GMM-based methods offer the advantage of being general and can be used to model many types of audio signals. These methods are called *general* or *a priori* models, as they are based on models which cover the range of properties for specific sources [20].

In this paper, we use a model-based approach based on vector quantization to separate audio sources according to their observed mixture recorded by single microphone. We employ a maximum likelihood (ML) amplitude estimator as our mixture estimator to find the two best states each from one codebook that when mixed best represent the current frame of the observed mixed audio signal. These states are then passed to a reconstruction stage (overlap and add procedure) in order to produce the separated output signals for each source. Through different experiments, we show that the proposed separation approach shows certain improvement compared to other mixture estimators including PSD [16],[17] and log-max [21].

The rest of the paper is organized as follows: In the following section, we present a review on existing state-of-the-art methods for single-channel audio separation. In section 3, the proposed model-based approach is presented. Later in this Section, we derive an ML mixture estimator which is used to find the best states of the underlying signal models. We present new concept of non-linear mask and replace common linear masks with these new masks. Section 4 presents the experimental results. Section 5 concludes on the work.

II. MODEL-BASED SINGL-CHANNEL AUDIO SOURCE SEPARATION

In this section we formulate the single-channel audio source separation problem. Consider an audio mixture denoted as $x(n)$ formed by speech and music signals as

$$x(n) = m(n) + s(n) \qquad n = 1, \ldots, N \quad , \tag{1}$$

where $m(n)$ is the music signal and $s(n)$ is the speech signal, N is the time window length in sample and n is the time sample index. The objective of a model-based approach for single-channel audio separation algorithm is to use the pre-trained statistical models to recover the unknown audio signals from their observed mixture. Using the additivity property of the spectral components in the short-Time *Fourier* Transform (STFT) representation, the mixed signal is given by

$$X(k) = M(k) + S(k) \quad , \tag{2}$$

where k denotes the frequency index and $X(k)$, $M(k)$ and $S(k)$ are the frequency domain representation for the mixed audio, speech and music signals, respectively. In [17], diagonal covariance matrices were used in the statistical models for both speech and music sources. According to [17], at each frame, the resulting Bayesian estimation for the separated audio sources is obtained as a Wiener-like filter defined as below,

$$\hat{M}(k) = \frac{\sigma_M^2(k)}{\sigma_M^2(k) + \sigma_S^2(k)} X(k) \quad , \tag{3}$$

$$\hat{S}(k) = \frac{\sigma_S^2(k)}{\sigma_M^2(k) + \sigma_S^2(k)} X(k) \quad , \tag{4}$$

where $\hat{M}(k)$ and $\hat{S}(k)$ are the estimated music and speech signals after separation, $\sigma_M^2(k), \sigma_S^2(k)$ are the diagonal elements in the covariance matrix related to music and speech signals, respectively. According to [17], using a mask-based approach as shown in (3) and (4), could introduce some undesired cross-talk from the interfering source signal into the separated signal. In particular, the resulting separated signal obtained by GMM-based audio separation algorithms with Wiener filter often suffer from the affects of the other source signal and the separation performance can be degraded in terms of the perceived audio [7]. This is caused by the components of the interference signal which remain in the separated output of the desired source signal.

III. The Proposed Separation Method

A. Feature extraction and codebooks

According to [22], the selected feature parameters together with the statistical models trained for the underlying signals in the mixture play the key role and determine the resulting performance for audio enhancement. As our feature parameter, we use the STFT spectrum amplitude of the audio and speech sources. This is comparable to the previous separation methods which used logarithm of the spectrum as their selected feature. As our statistical model, we train VQ codebooks for both music and speech from comprehensive datasets by using a modified VQ method described in Appendix I.

B. Proposed mixture estimator

In this section, we present the mathematical analysis to derive new mixture estimator based on the STFT features for single-channel audio source separation. The observed mixed signal, *X*(*k*) depends on both amplitude and phase of the underlying discrete Fourier transform (DFT) spectra. Considering $\theta_M(k)$ and $\theta_S(k)$ as the phase values of the speech and music signals, the mixed signal is given by $M(k)e^{j\theta_M(k)}+S(k)e^{j\theta_S(k)}$ and we have

$$X(k)=\sqrt{M^2(k)+S^2(k)+2M(k)S(k)\cos\theta(k)} \quad , \qquad (5)$$

where *θ(k)* is the phase difference defined as $\theta(k)=\theta_M(k)-\theta_S(k)$. It is already well-known that presenting a compact model for phase values is a difficult task; hence, in order to exclude the phase information in current problem, it is required to have a mixture estimator independent of phase information. In [21] it is shown that log-max approximation of the log spectrum of the mixed signal (here *x*(*n*)) is very close to the element-wise maximum of the log spectra of the two underlying signals (here *m*(*n*) and *s*(*n*)). The audio source separation problem presented in (1) can be considered as the problem of finding estimate of *X(k)* and not its logarithm used in [21] in a mean square error (MSE) sense and we have [23],

$$\begin{aligned}\log X_{MSE}(k) &= E\{\log X(k)\,|\,\log S(k),\log M(k)\}\\ &=\max\{\log S(k),\log M(k)\}\end{aligned} \quad , \qquad (6)$$

with *k*=1,2,…,*D* as the frequency index and *D* as the number of DFT-points used. In the following, we derive an ML mixture estimator in the DFT-domain. We find the probability distribution function (PDF) of the mixed signal, *X*(*k*) as a random variable. We derive an ML mixture estimator for *X*(*k*) based on the audio unknown signals in the mixture i.e. *S*(*k*) and *M*(*k*). The ML estimation can be considered as

$$\hat{X}(k)=\max\, p_{X(k)}\left(X(k)\,|\,S(k),M(k)\right) \quad . \qquad (7)$$

To derive the ML-estimator, one needs to calculate the PDF for the mixed signal given by $p_{X(k)}\left(X(k)\,|\,S(k),M(k)\right)$. In order to sake the simplicity in our notations, hereafter, we let as follows $p_{X(k)}(X(k))=p_{X(k)}\left(X(k)\,|\,S(k),M(k)\right)$. It is already shown that for audio signals the phase distribution can be well approximated by a uniform distribution [21],[24]. Considering a uniform phase distribution for both speech and music, then we have $p_s(\theta_s(k))=1/2\pi$ and $p_m(\theta_m(k))=1/2\pi$, $-\pi\le\theta_s(k)\le\pi$. As a consequence, the distribution of the mixture phase $\theta(k)$ is $p_{\theta(k)}(\theta(k))=p_{\theta_m(k)}(\theta_m(k))*p_{\theta_s(k)}(\theta_s(k))$ where * denotes the convolution operator. The PDF for the mixed signal $p_{\theta(k)}(\theta(k))$, is then obtained as

$$\begin{aligned}p_{\theta(k)}(\theta) &= p_{\theta_m(k)}(\theta_m(k))*p_{\theta_s(k)}(\theta_s(k))\\ &=\begin{cases}\dfrac{1}{4\pi^2}(\theta(k)+2\pi), & -2\pi\le\theta(k)\le 0\\ -\dfrac{1}{4\pi^2}(\theta(k)-2\pi), & 0\le\theta(k)\le 2\pi\end{cases}\end{aligned} \quad . \qquad (8)$$

Considering the periodicity of phase, then we obtain

$$p_{\theta(k)}(\theta(k))=\frac{1}{2\pi} \qquad -\pi\le\theta(k)\le\pi \quad . \qquad (9)$$

Given the distribution for mixture phase, *θ*(*k*), the PDF of *X*(*k*) can be obtained by the following formula [23],

$$p_{X(k)}(X(k))=\frac{p_{\theta(k)}(\theta(k))}{\left|\dfrac{\partial X(k)}{\partial\theta(k)}\right|} \quad . \qquad (10)$$

Using (5) and (10) together then we obtain

$$\frac{\partial X(k)}{\partial\theta(k)}=\frac{-S(k)M(k)\sin\theta(k)}{\sqrt{S^2(k)+M^2(k)+2S(k)M(k)\cos\theta(k)}} \quad . \qquad (11)$$

Finding the equivalent from for term $\sin\theta(k)$ in (5), and then plugging the result into (11) we have

$$P_{X(k)}(X(k)) = \frac{X(k)}{2\pi M(k)S(k)\sqrt{1-\left(\frac{X(k)^2-S(k)^2-M(k)^2}{2M(k)S(k)}\right)^2}} \quad , (12)$$

and $\left|S(k)e^{j\theta_s(k)} - M(k)e^{j\theta_m(k)}\right| \le X(k) \le \left|S(k)e^{j\theta_s(k)} + M(k)e^{j\theta_m(k)}\right|$ i.e. the audio mixture, $X(k)$ is limited by two extreme cases: 1) $\left|S(k)e^{j\theta_S(k)} - M(k)e^{j\theta_M(k)}\right| \approx 0$ when the DFT amplitude of the sources are too close to each other, and 2) when one of the sources dominates the other i.e. $S(k) >> M(k)$ or $M(k) >> S(k)$ then we have $\left|S(k)e^{j\theta_S(k)} - M(k)e^{j\theta_M(k)}\right| \approx S(k)$ and $M(k)$, respectively. Fig. 1 depicts the PDF given by (12). As it is observed from these plots, $S(k)+M(k)$ can be considered as the ML mixture estimate since its occurrence is most probable for either of the two extreme cases mentioned above. The ML mixture estimator is expressed as below

$$\hat{X}_{MLA}(k) = \max p_{X(k)}(X(k)) = S(k) + M(k) \quad , \quad (13)$$

where |.| denotes absolute value for frequency bin k. To verify the mixture estimator in (13) in practice, we calculated the distribution of the phase information of the mixed signal $X(k)$ using 100,000 frames extracted from audio signals selected from the database (see Section 4).

The audio mixtures we used are composed of speech and music signal where music signals are selected as solo and piano and speech signals are selected from our database. Fig. 2 shows the results as histogram plots for two extreme cases: 1) when the DFT spectrum amplitude are too close, and 2) when one of the underlying signals is dominant than the other. It is observed from these figures that the occurrence for the proposed mixture estimator is the most probable for both cases. This confirms that the mixture estimator is a ML-based estimator.

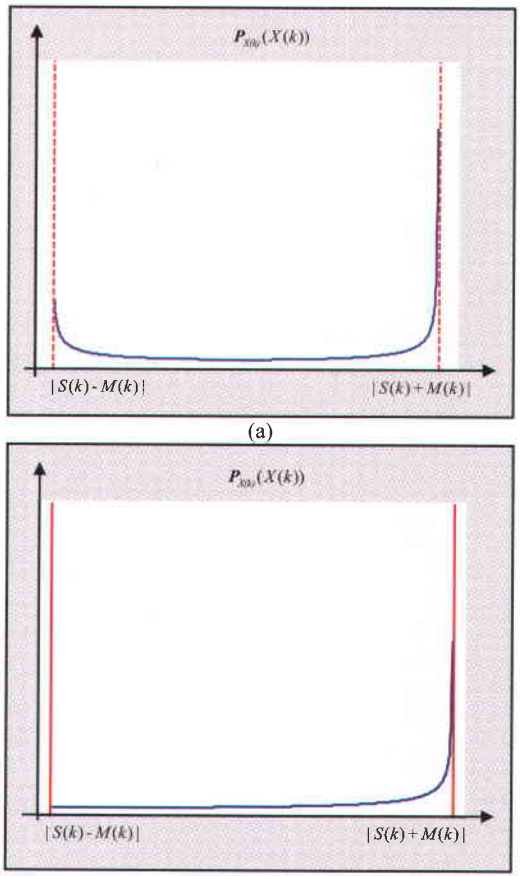
C. Separation stage

In order to find the best states of models of the underlying signals in the mixture, we employ the ML mixture estimator in (13). We search all possible states of the VQ codebooks of the signals to find those two indices which when combined minimize a distortion measure. Mathematically, these indices are found by solving the a minimization problem as

$$i_{opt}, j_{opt} = \underset{i,j}{\arg\min}\left\{X(k) - M_i(k) - S_j(k)\right\} \quad , \quad (14)$$

where $M_i(k)$ and $S_j(k)$ are codevectors related to the state ith and jth states corresponding to the VQ codebooks of the music and speech, respectively. The proposed mixture estimator in (14), is comparable to the previously proposed PSD method in [16],[17] as

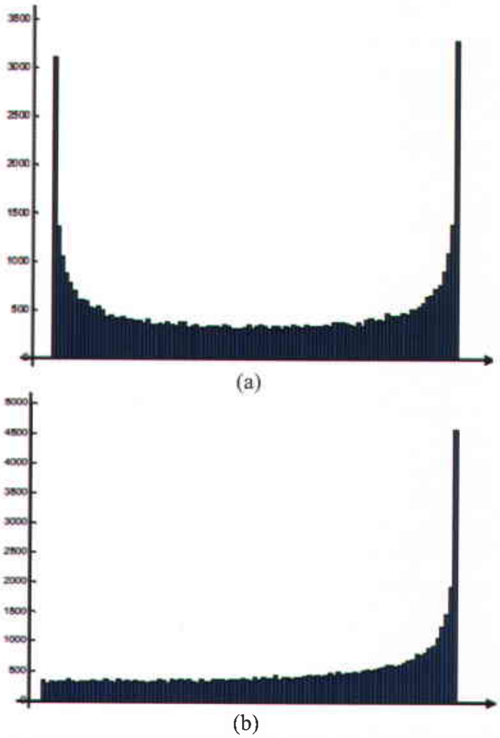
$$i_{opt}, j_{opt} = \underset{i,j}{\arg\min}\left\{X^2(k) - M_i^2(k) - S_j^2(k)\right\} \quad . \quad (15)$$



(a)

(b)

Fig. 1. Showing the PDFs of the mixed signal in (12), $X(k)$ for two extreme cases: (a) $|S(k)-M(k)| \approx S(k)$ or $M(k)$, and (b) the amplitude spectra of the signals are close to each other i.e. $|S(k)-M(k)| \approx 0$



(a)

(b)

Fig. 2. Showing the PDFs of the mixed signal in (12) and verifying it by calculating the histograms of the DFT amplitude for $X(k)$ for two scenarios: (a) $|S(k)-M(k)| \approx S(k)$ or $M(k)$, and (b) the amplitude spectra of the signals are close to each other i.e. $|S(k)-M(k)| \approx 0$

Another well-known mixture estimator is the conventionally used log-max approximation in [21] given as

$$i_{opt}, j_{opt} = \underset{i,j}{\arg\min}\left\{\log X(k) - \max\left(\log M_i(k), \log S_j(k)\right)\right\} \quad ,(16)$$

where i and j indicate the possible states for music and speech signal, while i_{opt} , j_{opt} denote the best states in the codebooks of the music and speech, respectively. The log-max approximation only considers the maximum element-wise of the logarithm spectrum magnitude of the signals in the mixture.

D. Using nonlinear masks

According to our derivations in the separation stage, the ML-estimate for mixture is given by

$$\hat{X}(k) = M_{i_{opt}}(k) + S_{j_{opt}}(k) \quad . \qquad (17)$$

The states of the two sources are sent to the reconstruction stage and are used along with the mixture phase to re-synthesize the recovered audio signals using a Wiener type formula as

$$\begin{aligned} \hat{M}(k) &= Mask_1(k)X(k) = \frac{M_{i_{opt}}(k)}{M_{i_{opt}}(k)+S_{j_{opt}}(k)}X(k) \\ \hat{S}(k) &= Mask_2(k)X(k) = \frac{S_{j_{opt}}(k)}{M_{i_{opt}}(k)+S_{j_{opt}}(k)}X(k) \end{aligned} \quad ,(18)$$

where $Mask_1$ and $Mask_2$ indicate the masks to recover music and speech signals, respectively. These estimates are then used along with the mixture phase in order to reconstruct the separated signals as

$$\begin{aligned} \hat{s}(n) &= DFT^{-1}\left\{\hat{S}(k)e^{j\angle X(k)}\right\} \\ \hat{m}(n) &= DFT^{-1}\left\{\hat{M}(k)e^{j\angle X(k)}\right\} \end{aligned} \quad . \qquad (19)$$

The linear mask in (18) causes large amount of undesirable crosstalk in the audio separation results. In order to solve the undesirable crosstalk problem occurred in linear mask (Wiener filtering), here, we propose a new mask called *nonlinear mask*. In the proposed mask the frequency bins whose values are higher than 0.5 are retained the same as they are using a linear transfer function. As a result in this region we block any interference to be introduced in the separated output signal. Such linear function ranges between [0.5,1] and would result in a better perceptual quality in the separated output signal. On the other hand, the values lower than 0.5 implicitly state that the second audio source has higher frequency amplitude. As a result, it is reasonable to set the audio source with smaller amplitude to zero. This will guarantee to remove crosstalk effects already existed in separated audio signals in previous masks. We call this new mask as semi-soft mask which result in the high perceptual signal quality by setting crosstalk to its minimum value. Fig. 3 demonstrates the semi-soft mask.

In general, we propose non-linear mask defined as follow; a non-linear mask emphasizes the spectral components with values higher than 0.5 while attenuating those with values smaller than 0.5. In particular, according to the definition of ideal binary mask we expect that a mask should satisfy the following constraint to establish a tradeoff between low crosstalk and low speech distortion. These constraints are defined as follow. Assume that x is the value of the mixture spectrum at *k*th frequency bin. Then it is required that the slope of the mask curve gets to zero at x=0 and x=1. Hence we have $\partial Mask_i(x=0)/\partial x=0$ and $\partial Mask_i(x=1)/\partial x=0$ where $i=\{1,2\}$ denoting each of the two underlying signals. The value of mask at x=0 and x=1 is required to satisfy some boundary constraints at x={0,1} requiring that Mask(x=0)=0 and Mask(x=1)=1. In addition, to have a symmetric mask, we need that the mask curve satisfy a symmetry property around x=0.5. Taking all these constraints into account, we can show that the function satisfying all these constrains is of the form

$$\text{Mask}^K(X) = \frac{1+\sqrt[K]{2X-1}}{2} \quad , \qquad (20)$$

whose values are determined based on the values of the desired signal $X(k)$. Equation (20) is considered as the general formulation for the family of non-linear masks. Fig. 4 illustrates some examples for curves for the proposed non-linear masks obtained by several values of K in (20). Applying the nonlinear mask in (20) into our audio separation problem, the masks required to recover the music and speech signal are found as below,

$$\begin{aligned} MASK_S^K(k) &= \frac{1+\sqrt[K]{2MASK_S(k)-1}}{2} \\ MASK_M^K(k) &= \frac{1+\sqrt[K]{2MASK_M(k)-1}}{2} \end{aligned} \quad . \qquad (21)$$

It is important to note that by employing different values for parameter K in (21) results into different separation scenarios described as below

$$\begin{cases} K=1 & :\textit{linear soft mask} \\ K=3,\ 5\ ,7 & :\textit{nonlinear soft mask} \\ K=\infty & :\textit{binary mask} \end{cases} \quad . \qquad (22)$$

Letting K=1, leads into linear mask which maximizes the cross-talk of the other source but the re-synthesized perceived signal quality for the separated output signal is rather acceptable. This is confirmed through our experiments as discussed in the following Section. This indicates that by using linear masks (K=1) in (18), the interference is audible along with the desired audio source in the separated signals. This degrades the separation performance and consequently would result in an inferior separation performance. As another extreme case, by letting

K=∞, the resulting mask is degenerated to the well-known binary mask where the crosstalk is maximum and the perceived signal quality of the underlying sources is significantly dropped but the signal distortion is low (high signal intelligibility). Here, by considering the nonlinear mask proposed in (20), we try to keep the advantages of both binary mask (K=∞) and linear mask (K=1) discussed above. In our experimental results we demonstrate that using nonlinear mask results in a lower signal to distortion ratio (SDR).

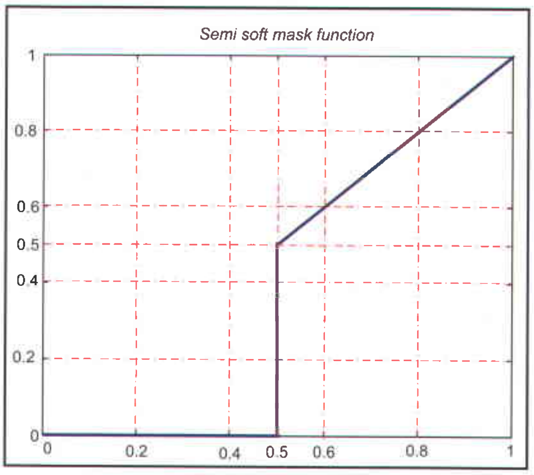


Fig. 3. Showing the proposed semi-soft mask in (20) vs. X(k) [25].

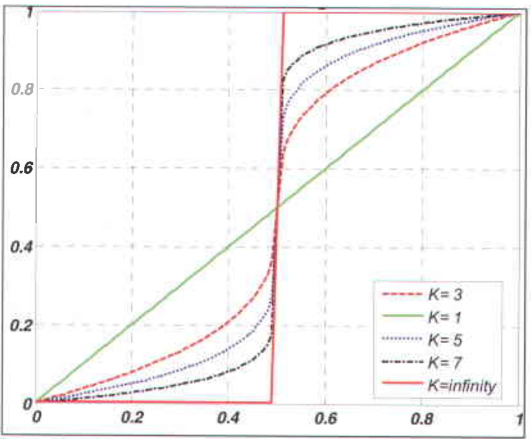


Fig. 4. Showing examples of curves for the non-linear mask in (20) evaluated for different values of K=1,3,5,7,∞ [25].

E. Separation scenario

The whole separation method is shown as a block diagram in Fig. 5. The audio mixture, $x(n)$ is entered to a DFT block. The result, $X(k)$ is entered to the ML mixture estimator in (14). Two states are found denoted by i_{opt} and j_{opt} for music and speech, respectively. These indices encode the magnitude spectrum vector in the VQ codebooks of the underlying signals in the mixture. These selected states in the codebooks along with the mixture phase are then passed to an inverse DFT (IDFT) block. The separated music and speech signals are reconstructed using an overlap-add procedure. In the reconstruction stage we use phase synchronization proposed in [26]. This is implemented by making phase values coherent at the frame boundaries also called phase coherency.

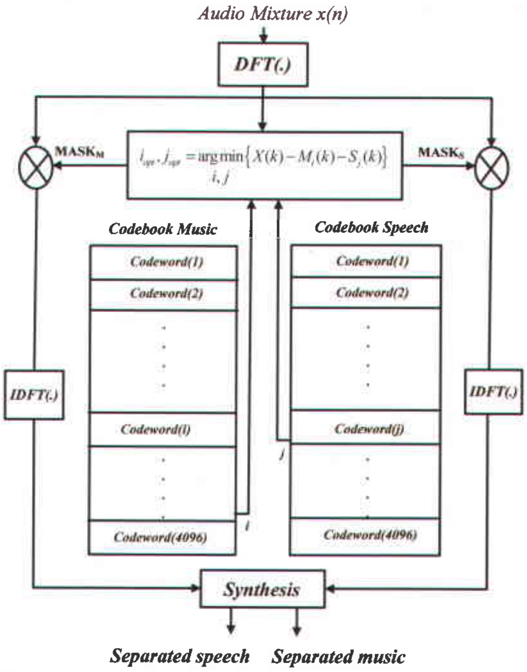


Fig. 5. The block diagram for the proposed method in single-channel audio separation.

IV. Simulation and Experimental Results

A. System setup and database

As a proof of concept, in this section we evaluate the separation performance obtained by the proposed method and compare it with other benchmark methods in audio source separation. As our speech database, we use a rather comprehensive database prepared by our laboratory members consisting of 100,000 sentences uttered by more than 25 speakers including 15 male and 10 female speakers. The sentences to be uttered by the speakers are collected from 230 phonetically balanced sentences. Each speaker uttered one sentence in five different ways differing in the stress point. As our music database we collected 2 compact discs consisting of 2 hours of piano and solo audio tracks. The number of vectors extracted from the audio files is 200,000. These vectors are then used to produce codebooks as signal models for the signals in the mixture. We trained speaker dependent codebooks for each speaker male/female and music.

As shown in Fig. 5, the core of the separation system is composed of two VQ codebooks: one for speech source and the other for music source. The mixture estimation stage finds the indices denoted by i_{opt} and j_{opt} as the optimal indices for speech and music codebooks, respectively. These indices are then used to produce nonlinear masks proposed in this paper. These masks are then applied to the mixed signal. The filtered signals are then used along with the mixture phase to recover the time domain separated output signals for speech and music signals.

As our settings for separation, we use an analysis window of 32 msec with a frame shift of 16 msec.

The sampling frequency is set to 8 kHz. The codebook size is 12 bit (with 4096 codeword entries). Note, the scenario taken here is more difficult than the scenarios already reported in [5],[8],[10],[16] where the audio mixtures are only composed of music tones and vocal speech frames which are highly harmonic. In this work, we consider a more challenging separation problem where speech and music signals are highly overlapped in their spectrum.

In order to evaluate the separation performance of the proposed method, we use SDR as our objective measure proposed in [27]. The SDR is considered as a measure to ensure how much interfering signal is separated from the mixture. The SDR measure is calculated in the frequency domain defined as

$$\mathrm{SDR} = 10\log_{10}\left[\frac{\sum_k X^2(k)}{\sum_k\left(X(k)-\hat{X}(k)\right)^2}\right] , \qquad (23)$$

where $X(k)$ and $\hat{X}(k)$ are the DFT spectrum for the original and the reconstructed signals, respectively. The results are reported after averaging the SDR measures obtained from different pairs of speech + music signals mixed at 0 dB level. Listening experiments were conducted. A total of 7 persons of various ages participated and were trained for the test. The subjects included 5 women and 5 men with graduate-level educations. The participants listened to the original and synthesized signals by the proposed method and other benchmark methods (PSD and Log-max). Then they were asked to give an opinion score from 1 to 5 (where 1=bad and 5=excellent quality). The clips played for the participants consisted of ten clips composed of speech + music mixed signals. The MOS was obtained by averaging the results. The results of the mean opinion score (MOS) are reported in Tables 3 and 5 as our subjective evaluation.

B. Results

Table 1 and 2 summarizes the results for the averaged SDR for two different mixing scenarios: music signals are mixed with 1) male speaker, and 2) female speaker. The results are reported for the proposed method in this paper and benchmark methods: log-max mixture estimator in (16) [21], and the PSD method given in (15) used in [16],[17]. It is observed that, the averaged SDR results for the proposed method outperforms log-max by 1 dB and PSD by 2 dB. Table 3 demonstrates the obtained MOS score for various estimators including the proposed method, log-max and PSD. In addition, the effect of employing phase synchronization is also explored in the Table 3. It is observed that the phase synchronized separation output signals advantageously lead to significantly more acceptable perceived signal quality as indicated by the MOS results. The significant difference between the results in Table 2 and 3 confirms that the SDR measure may not fully reflect the real perceived quality of a separation scenario which is in agreement with [27].

Table 4 shows the averaged SDR results for different masks. We included linear mask (with K=1), binary mask (with $K=\infty$) and other possible values of K=3,5,7. From the results shown in Table 4, it is concluded that selecting K=3 in the proposed nonlinear mask results in the least SDR. On the other hand, the binary mask results in the poorest performance as indicated by the SDR values. Fig. 6 depicts the mixture of speech and music as well as the accordingly their separated outputs both in a time and a spectrogram representation.

As our subjective results, we use the MOS results to evaluate the separated signals while using different masks to compare the performance of our proposed nonlinear masks with both linear mask and binary mask. Table 5 shows the MOS results for different masks. From the results in Table 5, it is observed that the proposed non-linear masks outperforms others.

Table 1: Averaged SDR in (dB) for separated male speech and music signals using the proposed, log-max and PSD algorithms.

Signal	PSD	Log-max	Proposed
Speech(male)	5.44	6.23	8.51
Music	9.76	11.74	12.24

Table 2: Averaged SDR for separated female speech and music signals using the proposed, log-max and PSD algorithms.

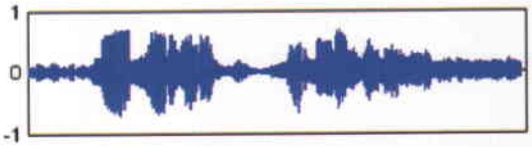
Signal	PSD	Log-max	Proposed
Speech(female)	7.61	9.40	9.91
Music	10.39	12.09	12.03

Table 3: MOS results for the output signals obtained by different separation methods.

Method	Scenario	Synthesis Method	MOS
Proposed	Male	Phase Synchronization	4.3
		No Phase Synchronization	3.7
	Female	Phase Synchronization	3.2
		No Phase Synchronization	2.8
Log-max	Male	Phase Synchronization	2.5
		No Phase Synchronization	2
	Female	Phase Synchronization	2.4
		No Phase Synchronization	1.9
PSD	Male	Phase Synchronization	2.2
		no Phase Synchronization	1.7
	Female	Phase Synchronization	2.1
		no Phase Synchronization	1.6

Table 4: Averaged SDR for separated speech and music in masking approach

Method	Category	SDR
Linear mask	Speech	7.88
	Music	18.77
nonlinear mask (K=3)	Speech	8.27
	Music	19.55
nonlinear mask (K=5)	Speech	8.15
	Music	19.50
nonlinear mask (K=7)	Speech	8.04
	Music	19.42
binary mask (K=∞)	Speech	7.06
	Music	17.91



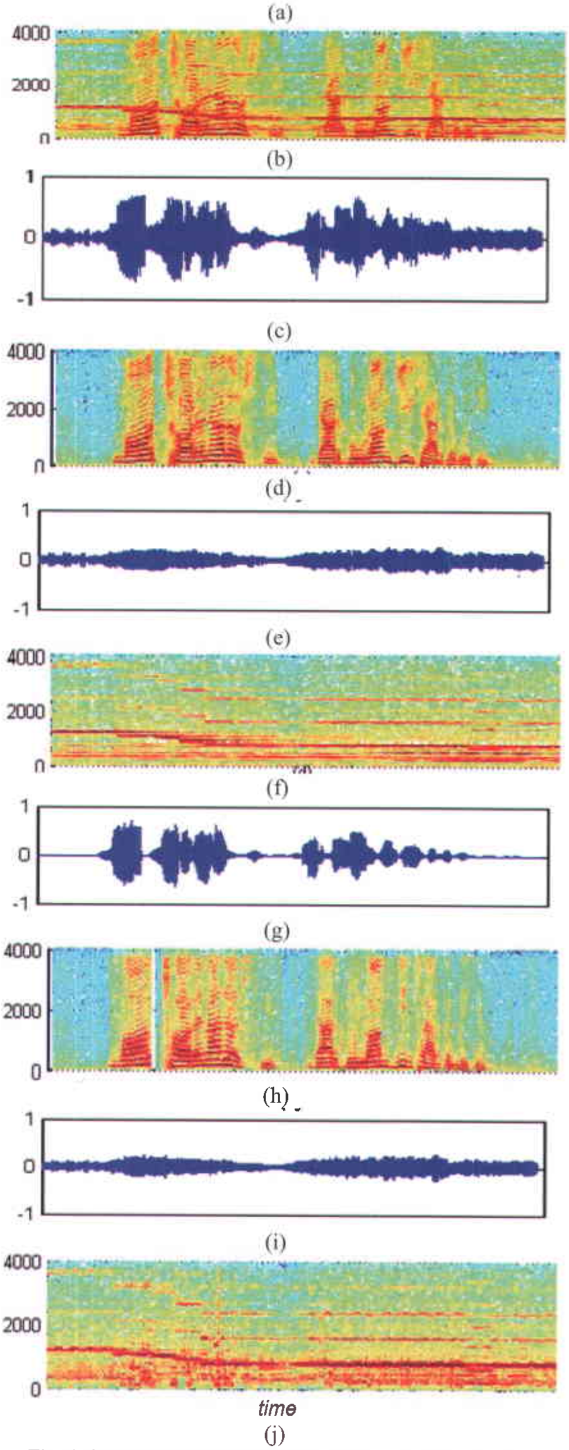


Fig. 6. Corresponding spectrograms and time domain of (a,b) mixed signal composed of speech+music, (c,d) speech, (e,f) music, (g,h) separated speech, (i,j) separated music signal.

Table 5: MOS results for output signals using mask.

Method	Category	MOS
Linear mask	male + music	2.3
	female + music	2.5
nonlinear mask (K=3)	male + music	3.5
	female + music	3.4
binary mask	male + music	1.7
	female + music	1.8

V. CONCLUSION

In this paper a low complexity mode-based audio source separation approach was proposed. The separation method was based on VQ codebooks trained on speech and music signals. We derived a maximum likelihood (ML) mixture estimator based on the spectrum amplitude of the short-time Fourier transform (STFT). We also proposed new nonlinear masks which could establish a tradeoff between lower crosstalk and high quality in the separated audio signals. Through extensive simulation and experiments we compared the separation performance of the proposed method with log-max and power spectral density (PSD) approach as our benchmarks. It was observed that the proposed method outperforms in terms of signal-to-distortion ratio (SDR). The method also attained higher perceived quality in its separated output signals. It was also observed that nonlinear mask resulted into a higher perceived signal quality as indicated by their high MOS results.

APPENDIX I: HISTOGRAM-BASED VECTOR QUANTIZATION

The method used for initialization stage of a vector quantization plays an important role in the overall quantization performance [9]. The problem lies in the fact that there is no guarantee to reach at the global minimum. In this appendix, we present a method to mitigate this problem and prevent the algorithm to be trapped in local minima during VQ centroid update.

In the following, consider that we have a large set of training vectors composed of frames from audio signals. Each vector is denoted by $\boldsymbol{T}_i$ where i represents the number of the training vector with $i \in [1,R]$ and R as the number of training vectors. As our initialization step, we select M reference vectors from the training set, $\boldsymbol{T}$. To this end, we scan all training vectors and calculate their Euclidean distance from each other. Then those training vectors having a distance lower than a pre-defined threshold denoted by δ are selected. This threshold is then used to determine which training vectors are close to each other. To select an appropriate value for δ we have tested the range of [0.01,0.1] and found that δ=0.05 results in an acceptable result. The procedure is repeated for all training vectors, and the repetition of each of these vectors are counted in a temporary vector called histogram index vector denoted by $\boldsymbol{H}_i$ in Table 6. By repeating the procedure for other vectors, the entries in the histogram vector are filled. Then after sorting this histogram vector in a decreasing manner, we select the first M indices. The related vectors to these indices are the most probable vectors. The *pseudo-code* for the proposed histogram-based initialization is shown in Table 6.

The Euclidean distance used in the VQ centroid update during the VQ design may not efficiently represent the closeness of two really similar vectors in the clustering process. We use a perceptual weighting shown in Fig. 7. Since the frequency range within [0,1 kHz] has the most perceptually importance, we put a fixed weighting and a exponentially decaying weight function is assigned to the rest of frequencies. As a consequence, the frequency components lying in the frequency range of higher than 1 kHz are deemphasized as shown in Fig. 7.

Table 6: The pseudo-code for the proposed histogram-based initialization for codebook design.

```
Step1:
$T_i$ ,   $i=1,2,.....,R$: Training set vector
$H_j$ ,   $j=1,2,.....,R$: Histogram vector
$M_k$ , $k=1,2,.....,R$: :Initial vector for VQ algorithm
Step 2:
$H_j = 1$ , $j=1,2,.....,R$
  for $i= 1:R-1$
    if $H_i \neq 0$
       for $j=(i+1):R$
         if dist $(T_i - T_j) \leq \delta$
         $H_i = H_i+1$
         $H_j = 0$
         end
       end
    end
end
Step 3:
 Sort histogram decreasingly and select first $M_k$
 $H = sort(H_j)$
 $M_{k=1:M} = T_i$ ,    $i \in index(H_s)$
```

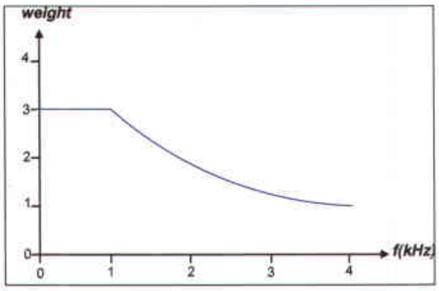


Fig. 7. Weighted distance employed in VQ algorithm.

REFERENCES

[1] 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.

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

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

[4] 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.

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

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

[7] G. J. Jang and T.-W. Lee, "A probabilistic approach to single channel source separation," Advances in Neural Information Processing Systems 15, MIT Press, Cambridge, 2003

[8] 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.

[9] T. Virtanen, "Unsupervised Learning Methods for Source Separation in Monaural Music," in *Signal Processing Methods for Music Transcription*, Springer-Verlag, pp. 267–296, 2006.

[10] T. Virtanen, "*Monaural Sound Source Separation by Non-Negative Matrix Factorization with Temporal Continuity and Sparseness Criteria*," IEEE Transactions on Audio, Speech, and Language Processing, vol 15, no. 3, March 2007.

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

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

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

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

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

[16] C. Fevotte , S. J. Godsill, "A Bayesian approach for blind separation of sparse sources", IEEE Trans. Speech Audio Process, vol. 4, no 99, pp. 1–15, 2005.

[17] L. Benaroya, F. Bimbot, and R. Gribonval., "Audio source separation with a single sensor," IEEE Trans. Audio, Speech and Language Processing, vol. 14, no. 1, pp. 191–199, Jan. 2006.

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

[19] A. Gersho and R. M. Gray, Vector quantization and signal compression, Kluwer Academic, Norwell MA, 1992.

[20] A. Ozerov, P. Philippe, F. Bimbot and R. Gribonval, "Adaptation of Bayesian models for single channel source separation and its application to voice / music separation in popular songs," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 15, no. 5, pp. 1564-1578, July 2007.

[21] 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.

[22] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[23] A. Papoulis, Probability, random variables, and stochastic processes, McGraw-Hill, 1991.

[24] 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.

[25] P. Mowlaee, A. Sayadian, M. Sheikhan, M. Fallah, "Single-channel music/speech separation using non-linear masks,"

International Symposium on Telecommunications (IST), pp. 543-547, Aug, 2008.

[26] P. Mowlaee, A. Sayadian, "A Fixed Dimension Modified Sinusoid Model (FD-MSM) for Single Microphone Sound Separation", International Conference on Signal Processing and Communications (ICSPC'07), Dubai, pp. 1183-1186, Nov. 2007.

[27] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'03)*, pp. 763–768, Apr. 2003.

Pejman Mowlaee was born in Anzali (Iran) in March 25th 1983, located at the southern cost of the Caspian Sea. He completed his Electrical Engineering (B.Sc.) with straight honors from Guilan University in 2005, and his Telecommunication Engineering in signal processing (M.Sc.) with straight honors from Iran university of Science and Technology (IUST) in 2007. He was a research assistant in the information processing research Laboratory (IPRL). Pejman received several awards, during his academic career so far, some of them as follow, Young researcher's Award for his M.Sc. study in 2006, Selected as a talent student with top class at Amirkabir university of Technology in 2008 and 2009, Honored M.Sc. thesis in the national wide contest for electrical students in the country (NSOEE) in 2007. His research interests include digital signal processing theory and methods with application to speech and audio, in particular single-channel separation, speech enhancement, speech coding, speech synthesis and microphone array processing. (Noise reduction, echo cancellation and DOA estimation). Currently, he is a research assistant at Multimedia and information signal processing (MISP) at Aalborg University.

Abolghasem Sayadiyan received the B.S. degree in Electrical Engineering from the Amirkabir University of Technology (Tehran Polytechnique), Tehran, Iran, the M.S. degree in Electrical Engineering from Esfahan University of Technology, and the Ph.D. degree in Electrical Engineering from the Department of Electrical and Computer Engineering, Tarbiat Modarres University, Tehran, Iran in 1980, 1987, and 1994, respectively. In September 1991, he joined the Department of Electrical Engineering, Amirkabir University of Technology (Tehran Polytechnique), Tehran, Iran where he is currently an Associate Professor. His research interest is in the area of speech processing. Dr. Sayadiyan is one of the pioneering researchers who have been developing Persian speech recognition algorithms and low bit-rate speech coders.

Hamid Sheikhzadeh Nadjar received the B.S. and M.S. degrees in electrical engineering from Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 1986 and 1989, respectively. He got his Ph.D. degree in electrical engineering from the Department of Electrical and Computer Engineering, University of Waterloo, Canada where he was also Research and Teaching Assistant. His research interests include signal processing and speech processing, with particular emphasis on speech recognition, speech enhancement, auditory modeling, adaptive signal processing, subband-based approaches, and algorithms for low-power DSP. He is a Senior Member of IEEE. Currently he is a faculty member at the Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran.