I.J. Information Technology and Computer Science, **2015, 03, 42-47**
Published Online February 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2015.03.06

Single Channel Speech Separation Using an Efficient Model-based Method

Sonay Kammi, Mohammad Reza Karami
Faculty of Electrical and Computer Engineering, Babol University of Technology, Babol, Iran
Email: sonaykammi@yahoo.com, mkarami@nit.ac.ir

Abstract— The subject of extracting multiple speech signals from a single mixed recording, which is referred to single channel speech separation, has received considerable attention in recent years and many model-based techniques have been proposed. A major problem of most of these systems is their inability to deal with the situation in which the signals are combined at different levels of energies because they assume that the data used in the test and training phase have equal levels of energies, where, this assumption hardly occurs in reality. Our proposed method based on MIXMAX approximation and sub-section vector quantization (VQ) is an attempt to overcome this limitation. The proposed technique is compared with a technique in which a gain adapted minimum mean square error estimator is derived to estimate the separated signals. Through experiments we show that our proposed method outperforms this method in terms of SNR results and also reduces computational complexity.

Index Terms— Single Channel Speech Separation, Vector Quantization, MIXMAX Approximation, Gain Estimation, Source Estimation

I. Introduction

Speech signals are seldom available in pure form for speech processing applications, and are often corrupted by acoustic interference like background noise, distortion, simultaneous speech from another speaker etc. In such scenarios, it becomes necessary to first separate the speech from the background. In particular, the task of separating overlapping speech from multiple speakers, called speech separation, is especially challenging since it involves separating signals having very similar statistic and acoustic characteristics. The separation problem has attracted immense research effort in the past two decades, more so for the case when the mixture is available only from a single channel and multi-channel approaches cannot be used. This single channel situation is called the single channel speech separation problem and the two speaker case can be formulated as $z(t) = x(t) + y(t)$, where $x(t)$ is the speech signal of speaker one and $y(t)$ is the speech signal of speaker two. Many techniques have been proposed to solve this problem. These approaches are mainly divided into two categories: source driven [1-4] and model-based methods [9-15].

As a major example for the first group, computational auditory scene analysis (CASA) has widely been studied [1]. Generally speaking, CASA-based methods aim at segregating audio sources based on possible intrinsic perceptual acoustic cues from speech signals [2]. For CASA systems, a reliable multi-pitch tracking component is critical to find pitch trajectories of two interfering speech signals [5]. The CASA methods are fast and could be implemented in real time. There are, however, challenges that limit the pitch tracking performance for a mixture [2]: (1) Most existing pitch estimation methods perform reliably only with clean speech signals that have a single pitch track or harmonically related sinusoids [6] with almost no background interference [7]. (2) It is possible to perform a reliable pitch estimation using a mixture of a dominant (target) and a weaker (masking) signal as long as the pitches of the masking and target speech are different in a short frame [3]. A high similarity between the interference and target pitch trajectories results in performance degradation of CASA methods [8]. (3) Because of energetic masking defined in [8], the weaker signal frames are masked by the stronger ones complicating the pitch estimation. Accordingly, at target-dominant time-segments, it is possible to accurately track the pitch contour of only the dominant (target) signal. (4) Pitch tracking performance has not been promising for scenarios where the underlying signals include mixtures of unvoiced and voiced frames and as a result, the separated speech signals include severe cross-talks [4].

Model-based single channel speech separation is commonly referred to as the techniques which use the trained models of the individual speakers to separate the sources from a single recording of their additive mixture. The most prominent models are vector quantization (VQ) [9], [13], Gaussian mixture models (GMM) [11], [12] and Hidden Marcov models (HMM) [14]. Given the individual speakers' models, an estimation technique is applied to estimate the sources. In most recent proposed model-based single channel speech separation techniques, it is assumed that the test speech files are recorded at a condition similar to that of the training phase recording. This assumption is not, however, realistic and highly limits the usefulness of these techniques. Therefore, it is of great importance to consider situations in which the test speech files are mixed at an energy ratio different from that of the training speech files. In these situations, a desired technique is one that first estimates the gains associated with the individual speakers.

In [15], a technique is proposed in which, it is shown that gains of the speech signals can be expressed in terms of a signal-to-signal ratio (SSR) and using this relation a gain adapted minimum mean square error (MMSE) estimator is derived to estimate the sources. Following

that, the patterns of the speakers and SSR which best model the observed signal in an MMSE sense are obtained. However this method sounds efficient to gain estimation, but it results long time processing in practice. In our proposed method we take the superiority of VQ which is simplicity computation to separate the speech signals [13]. In this paper we introduce sub-section VQ and use it instead of conventional VQ to achieve high accuracy in estimating gains and speech signals.

The rest of this paper is organized as follows. In section II, preliminary definitions are described where we express the sources-observation relation in the feature space and also the relation between speakers' gains and energies of the underlying signals. In section III, we give a description of training phase. In this section we introduce sub-section VQ method and show how it is applied in separation process. In section IV, details are given on how the gains of the speakers are estimated and how the estimated gains are applied to estimate the sources. Experimental results are reported in section V where the proposed technique is compared with a gain adapted MMSE estimator [15]. Finally, conclusions are given in section VI.

II. Preliminary Definitions

A. Gain-SSR Relation

In gain adapted methods, the relation between observation signal and the two sources is supposed to be

$$z(t) = g_x x(t) + g_y y(t) \qquad t = 0,1,\dots,T-1 \tag{1}$$

where g_x and g_y, which are positive parameters, are speakers' gains and it's supposed that these speech signals have equal power before gain scaling, $G_0^2 = \frac{1}{T}\sum_t x^2(t) = \frac{1}{T}\sum_t y^2(t)$. In [15] the speakers' gains are obtained in terms of SSR (signal to signal ratio), square root of power of the observation signal and G_0^2

$$g_x \approx \frac{g_z}{G_0\sqrt{1+10^{\frac{-SSR}{10}}}} \quad \text{and} \quad g_y \approx \frac{g_z}{G_0\sqrt{1+10^{\frac{SSR}{10}}}} \tag{2}$$

where $g_z^2 = \frac{1}{T}\sum_t z^2(t)$ is power of the observation signal and SSR=$10\log_{10}\frac{g_x^2}{g_y^2}$. Also, a_x and a_y are defined as

$$a_x = \log_{10} g_x \quad \text{and} \quad a_y = \log_{10} g_y \tag{3}$$

which will be used in gain estimation and source estimation process.

B. Sources-Observation Relation

Log magnitude of discrete furrier transform was selected as our feature. Let x(l) $l = 0,1,\dots,L-1$ be the samples of some speech signal segment (frame), possibly weighted by some window function, and let $X(e^{j2\pi d/L})$ denote the corresponding short time furrier transform.

$$X\left(e^{\frac{j2\pi d}{L}}\right) = \sum_{l=0}^{L-1} X(l) e^{-\frac{j2\pi ld}{L}} \qquad d = 0,1,\dots,D-1 \tag{4}$$

Let X denote the D dimensional, log spectral vector (feature vector) with dth component, X(d), defined by

$$X(d) = \log_{10}\left|X(e^{j2\pi d/L})\right| \qquad d = 0,1,\dots,D-1 \tag{5}$$

The relations between x[l], $\left|X(e^{j2\pi d/L})\right|$, $\angle X(e^{j2\pi d/L})$ and X(d) are shown in Fig. 1.

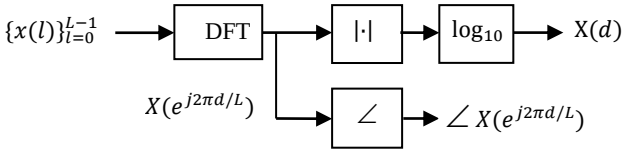

Fig. 1. Feature extraction

Let x^r and y^r be the L-dimensional vectors of the rth frames for the speech signals of speaker one and two in the time domain, respectively and z^r be the corresponding frame of the observation signal. We next form the following vectors as the feature vectors according to Fig. 1

$$X^r = \log_{10}(|F_D(x^r)|) \tag{6}$$

$$Y^r = \log_{10}(|F_D(y^r)|) \tag{7}$$

$$Z^r = \log_{10}(|F_D(z^r)|) \tag{8}$$

where X^r, Y^r, and Z^r denote the D-dimensional log spectral vectors of speaker one, speaker two, and the mixed signal, $F_D(\cdot)$ denotes the D-point discrete Fourier transform, and $|\cdot|$ denotes the magnitude operator.

The relation between feature vectors of the observation signal and the sources can be obtained using MIXMAX approximation [16]

$$\hat{Z}^r = \text{MIXMAX}(X^r + a_x, Y^r + a_y) = [\max(X^r(1) + a_x, Y^r(1) + a_y), \dots, \max(X^r(d) + a_x, Y^r(d) + a_y), \dots, \max(X^r(D) + a_x, Y^r(D) + a_y)]^T \tag{9}$$

This relation is used in the separation process to estimate the sources.

III. Modeling the Sources

A. VQ Modeling

VQ is referred to the techniques in which a set of available data vectors $\Phi = \{\varphi_m\}$, $m = 1,2,\dots,M$ are partitioned into a number of clusters V_n, $n = 1,2,\dots,N$ such that $\Phi = \bigcup_{n=1}^{N} V_n$ and $\bigcap_{n=1}^{N} V_n = \emptyset$. Every cluster V_n is represented by a vector called a codevector c_n and the set of all codevectors is called a codebook $C = \{c_n$, $n = 1,2,\dots,N\}$. In this paper we use LBG algorithm for VQ [17]. In this algorithm, clustering is performed in a way two optimality criteria which are:

nearest neighbor condition

$$V_n = \{\varphi_m : \|\varphi_m - c_n\|^2 < \|\varphi_m - c_{n'}\|^2, \ n \neq n'\} \tag{10}$$

and centroid condition

$$c_n = \frac{\sum_{\varphi_m \in V_n} \varphi_m}{\sum_{\varphi_m \in V_n} 1} \qquad n = 1,2,\dots,N \tag{11}$$

are met. In VQ modeling, a codebook is obtained for every speaker using training feature vectors of that speaker.

B. Sub-section VQ Modeling

In this method, training log spectral vectors for each speaker are divided into four sub-sections and for every sub-section, a VQ model is obtained. Also, log spectral vector of observation (Z^r) is divided into four sub-sections such that $Z^r = [Z_1^r; Z_2^r; Z_3^r; Z_4^r]$. We assume $X^r = [X_1^r; X_2^r; X_3^r; X_4^r]$ and $Y^r = [Y_1^r; Y_2^r; Y_3^r; Y_4^r]$ in which, X_k^r, $k = 1,2,3,4$ is the kth sub-section vector of X^r and also, Y_k^r, $k = 1,2,3,4$ is the kth sub-section vector of Y^r. According to (9) we have

$$\hat{Z}_k^r = \text{MIXMAX}\left(X_k^r + a_x, Y_k^r + a_y\right) \quad k = 1,2,3,4 \tag{12}$$

Sub-section VQ models are used to estimate the gains and the sources.

IV. Estimating the Sources

The single channel speech separation presented in this paper involves three stages. In the first stage, the speakers' gains are estimated. In the second stage, the estimated gains of the speakers are used to estimate the sub-section feature vectors of each speaker and in the third stage, time domain signal of each speaker is obtained from feature vectors of the associated speaker. Fig. 2 illustrates our proposed gain adapted single channel speech separation method. As the figure shows, Z^r indicates the feature vector of the observation signal, C_k^x, $k = 1,2,3,4$ is the codebook of speaker x in the kth sub-section, C_k^y, $k = 1,2,3,4$ is the codebook of speaker y in the kth sub-section and $\hat{X}$ is estimated feature vector of speaker x.

A. Gain Estimation

As the Fig. 2 shows, the gains of the speakers are estimated in the first sub-section. In order to estimate the gains, we estimate the SSR. For a given SSR, a_x and a_y are obtained using (2) and (3). Let $\tilde{c}_{1_{i,j}}^{mix}$ be MIXMAX estimator of two arbitrary codevectors in the first sub-section, that is,

$$\tilde{c}_{1_{i,j}}^{mix} = \text{MIXMAX}\left(c_{1_i}^x + a_x, c_{1_j}^y + a_y\right) \tag{13}$$

where $c_{1_i}^x$ and $c_{1_j}^y$ are two arbitrary codevectors of speaker x and speaker y, respectively. All pairs of codevectors $\{c_{1_i}^x, c_{1_j}^y\}$ are compared to find the minimum mean square error (MMSE) compared to the observation signal's feature vector Z_1^r. SSR which minimizes the MMSE for all frames is selected as the estimated SSR

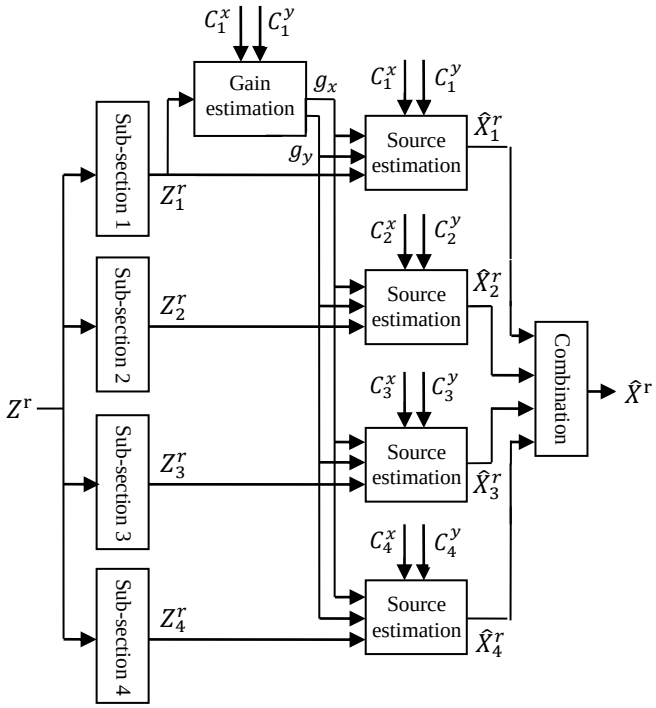

Fig. 2. Schematic of our proposed gain adapted single channel speech separation system

$$SSR^* = \arg\min_{SSR} \sum_r \min_{i,j} \left(\sum_d \left(Z_1^r(d) - \tilde{c}_{1_{i,j}}^{mix}(d) \right)^2 \right) \tag{14}$$

We define $Q(SSR) = \sum_r \min_{i,j} \left(\sum_d \left(Z_1^r(d) - \tilde{c}_{1_{i,j}}^{mix}(d) \right)^2 \right)$. Unlike the Q(SSR) defined in [15], it can be shown that our proposed Q(SSR) is a convex function. Here we give a brief description to prove it: using (13), (3) and (2), $Z_1^r(d) - \tilde{c}_{1_{i,j}}^{mix}(d)$ can be written as

$$Z_1^r(d) - \max\left(c_{1_i}^x(d) + \log_{10}\left(\frac{g_z}{G_0}\right) - \frac{1}{2}\log_{10}\left(1 + 10^{\frac{-SSR}{10}}\right), c_{1_j}^y(d) + \log_{10}\left(\frac{g_z}{G_0}\right) - \frac{1}{2}\log_{10}\left(1 + 10^{\frac{SSR}{10}}\right)\right).$$

Both $-\frac{1}{2}\log_{10}\left(1 + 10^{\frac{-SSR}{10}}\right)$ and $-\frac{1}{2}\log_{10}\left(1 + 10^{\frac{SSR}{10}}\right)$ are concave functions. So, both functions in max function argument are concave functions. Max of two concave functions is a concave function and when it's multiplied by a negative number, it becomes convex. So, $Z_1^r(d) - \tilde{c}_{1_{i,j}}^{mix}(d)$ is a convex function. Since power two of a convex function, summation of a number of convex functions and min of a number of convex functions are all convex functions, we deduce that $Q(SSR) = \sum_r \min_{i,j} \left(\sum_d \left(Z_1^r(d) - \tilde{c}_{1_{i,j}}^{mix}(d) \right)^2 \right)$ is a convex function. More information about convex functions can be found in [18].

As we proved above, Q(SSR) is a convex function, so it has a global minimum, mathematically $Q(SSR^*) < Q(SSR),\ \forall SSR \neq SSR^*$. To find SSR^*, we can use a very efficient iterative quadratic optimization algorithm presented in [19] and used in [15]. In this algorithm, three points are selected

$$\{(SSR_l, Q(SSR_l)), (SSR_c, Q(SSR_c)), (SSR_r, Q(SSR_r))\}$$

and they are updated each iteration to obtain a quadratic function of the form $f(x) = ax^2 + bx + c$. Update of the points is performed using $x^* = -\frac{b}{2a}$, the value that minimizes $f(x)$, and $Q(x^*)$. For initialization we regard: $SSR_l \leftarrow SSR_{min}$, $SSR_r \leftarrow SSR_{max}$, and $SSR_c \leftarrow$ an arbitrary value between SSR_{min} and SSR_{max}. The algorithm iterates until reaching a value of SSR_c that $Q(SSR_l) \geq Q(SSR_c) \leq Q(SSR_r)$. In experiments, we can see that for $SSR > 18$ dB the signal with higher energy which is called target signal, completely masks the signal with lower energy known as interference signal. So we can set $SSR_{min} = 0$ dB and $SSR_{max} = 18$ dB. Now, using the estimated SSR (SSR^*) we obtain the estimated gains of the sources (g_x^* and g_y^*) from (2). The estimated gains are used in the next subsection to estimate the feature vectors.

B. Source Estimation

Using the estimated gains of the sources, we obtain a_x^* and a_y^* from (3) and use them to form the MIXMAX estimator of two arbitrary codevectors of the speaker one and speaker two in the kth sub-section

$$\tilde{c}_{k_{i,j}}^{mix} = MIXMAX\left(c_{k_i}^x + a_x^*, c_{k_j}^y + a_y^*\right) \quad (15)$$

Then in each sub-section, for each frame we select the optimal codevectors that cause minimum mean square error (MMSE) between the feature vector of the observation signal and the MIXMAX estimator

$$\{i^*, j^*\} = \arg\min_{i,j}\left(\sum_d \left(Z_k^r(d) - \tilde{c}_{k_{i,j}}^{mix}(d)\right)^2\right) \quad (16)$$

Then, we use a simple soft mask filter to estimate log spectral vectors of speaker x and speaker y. In this method, the dth component of the estimated log spectral vector in the kth sub-section for speaker x is given by

$$\hat{X}_k^r(d) = \begin{cases} Z_k^r(d) - a_x^* & c_{k_{i^*}}^x(d) + a_x^* > c_{k_{j^*}}^y(d) + a_y^* \\ c_{k_{i^*}}^x(d) & c_{k_{i^*}}^x(d) + a_x^* < c_{k_{j^*}}^y(d) + a_y^* \end{cases} \quad (17)$$

Similarly, the estimation of $Y_k^r(d)$ is given by

$$\hat{Y}_k^r(d) = \begin{cases} Z_k^r(d) - a_y^* & c_{k_{j^*}}^y(d) + a_y^* > c_{k_{i^*}}^x(d) + a_x^* \\ c_{k_{j^*}}^y(d) & c_{k_{j^*}}^y(d) + a_y^* < c_{k_{i^*}}^x(d) + a_x^* \end{cases} \quad (18)$$

Then, the estimated sub-section vectors of each speaker are combined to obtain the entire feature vector of each speaker, $\hat{X}^r = [\hat{X}_1^r; \hat{X}_2^r; \hat{X}_3^r; \hat{X}_4^r]$, $\hat{Y}^r = [\hat{Y}_1^r; \hat{Y}_2^r; \hat{Y}_3^r; \hat{Y}_4^r]$.

C. Synthesizing Estimated Speech Signals

Here, the reverse of what we do for feature extraction is applied to the estimated log spectral vectors of each speaker to obtain time domain signals:

The estimated log spectral vectors of each frame ($\hat{X}^r$ and $\hat{Y}^r$) are transformed to the spectral domain and combined with the phase of the observed signal. Then, a D-point inverse DFT is applied to transform the vectors to the time domain. Mathematically, the procedure is expressed by

$$\hat{x}^r = F_D^{-1}\left(10^{\hat{X}^r} \exp\left[(-1)^{\frac{1}{2}} \angle F_D(z^r)\right]\right) \quad (19)$$

and

$$\hat{y}^r = F_D^{-1}\left(10^{\hat{Y}^r} \exp\left[(-1)^{\frac{1}{2}} \angle F_D(z^r)\right]\right) \quad (20)$$

where $\angle$ denotes the phase operator, $\exp[\cdot]$ denotes the exponential function, and $F_D^{-1}(\cdot)$ represents the D-point inverse Furrier transform and $\hat{x}$ and $\hat{y}$ are the estimated frames of signals in the time domain. Finally the inverse transformed vectors are multiplied by a Hann window and then the overlap-add method is used to recover the sources in the time domain.

V. Experimental Results

Our proposed method is evaluated in this section and it's compared with the method presented in [15] which is a gain adapted MMSE estimator. The database we use for our experiments is presented in [20]. In our experiments, the sampling rate of the signals is decreased to 8 kHz from the original 25 kHz. In the test phase, to obtain mixed signals with different SSRs we select 10 pairs of speech files randomly that are not used in the training phase and mix them at SSRs equal to 0, 6, 12 and 18 dB. In the mixed signal, the speech signal that has higher gain is the target signal and the signal with lower gain is the interference signal. In both training and test phases, frames of the speech files are obtained using a hamming window whose length is 50 ms and it's frame shift equals 20 ms. In the phase of reconstructing separated speech files, a Hann window is used in overlap add method. To obtain feature vectors of them, a 512 point discrete furrier transform is applied to them and after taking log magnitude of them and discarding their symmetric portions, 257 dimensional feature vectors are obtained. Afterward, we select 128 codevectors as the size of the codebook. In order to generate sub-section VQ models, training feature vectors of each speaker are divided into 4 sub-sections, 65-point log spectral vectors for first sub-section and 64-point log spectral vectors for other sub-sections, and a VQ model is obtained for every sub-section.

The similarity between the original signal and the estimated signal of speaker x is measured by signal-to-noise ratio (SNR) which is defined as follows

$$\mathrm{SNR_x} = 10\log_{10}\left[\frac{\sum_t (x(t))^2}{\sum_t (x(t)-\hat{x}(t))^2}\right] \quad (21)$$

where $x(t)$ and $\hat{x}(t)$ are the original and estimated speech signals respectively.

Our experiments include two stages: in the first stage, in order to show the effectiveness of our proposed VQ based gain estimation approach in the first sub-section, we select 20 pairs of speech files randomly and mix them at random integer SSRs within the interval [0, 18] dB. Then we estimate the SSR for each mixture using (14) and compare it with the actual SSR. We also keep tracking of the number of iterations performed in the quadratic optimization algorithm to reach estimated SSR (SSR^*). Our experimental results of this stage are presented in Table 1. This table includes 4 columns. The first column determines the number of the mixture. The second column gives us the actual SSR of the corresponding mixture. The third column gives us the estimated SSR of the corresponding mixture and the forth column reports the number of iterations at which the quadratic optimization algorithm is reached to the corresponding estimated SSR. As it can be seen from the table, our proposed gain estimation approach estimates the SSR with reasonable accuracy and these estimated SSRs are reached with only one or two iterations for most of the cases.

Table 1. Experimental results of our proposed gain estimation method performed on 20 randomly selected pairs of speech files and mixed at random integer SSRs within interval [0, 18] dB

mixture	SSR^{actual}	SSR^*	Itr
1	7	6.0204	1
2	12	11.0332	3
3	5	5.6115	1
4	1	1.5112	1
5	10	8.7727	2
6	9	8.5915	1
7	6	7.5366	2
8	2	2.2444	1
9	15	15.1206	2
10	4	3.4199	1
11	8	7.8029	2
12	13	12.5192	1
13	17	16.4497	1
14	2	2.8750	2
15	11	11.9618	3
16	5	4.4533	1
17	3	4.0130	1
18	9	8.1751	1
19	14	12.2161	3
20	7	5.8547	1

In the second stage of our experiments, we try to show the effectiveness of the entire gain adapted single channel speech separation system we proposed. For this purpose, we select 10 pairs of speech files randomly that are not included in the training phase and mix them at SSRs equal to 0, 6, 12 and 18 dB. Fig. 3 and Fig. 4 show averaged SNR results versus SSR for separated target and interference speech signals respectively, for: our proposed method (◊ line), the method presented in [15] which is a gain adapted MMSE estimator (□ line), our proposed method without gain adaptation (Δ line), and MMSE estimator (o line). As the figures show, our proposed method outperforms the gain adapted MMSE estimator for both target and interference signals and also it can obviously be seen that when gain estimation is not included in that methods, separation performance greatly degrades which signifies importance of gain estimation in model-based methods. Our proposed method has much lower computational complexity with respect to gain adapted MMSE estimator, because both gain estimation and source estimation phases in our method deal with fewer parameters.

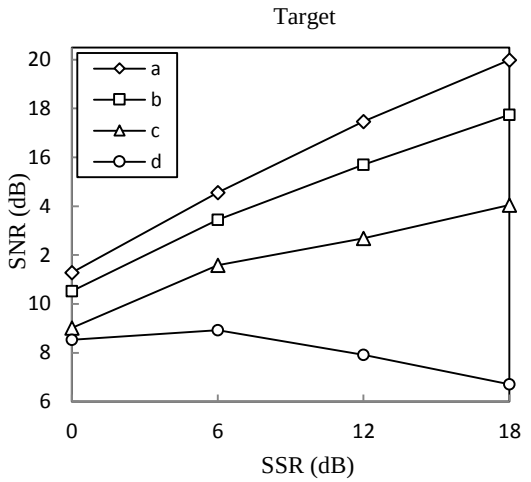

Fig. 3. averaged SNR versus SSR for separated target speech files obtained from our proposed method (a), gain adapted MMSE estimator (b), our proposed method without gain estimation (c) and MMSE estimator (d

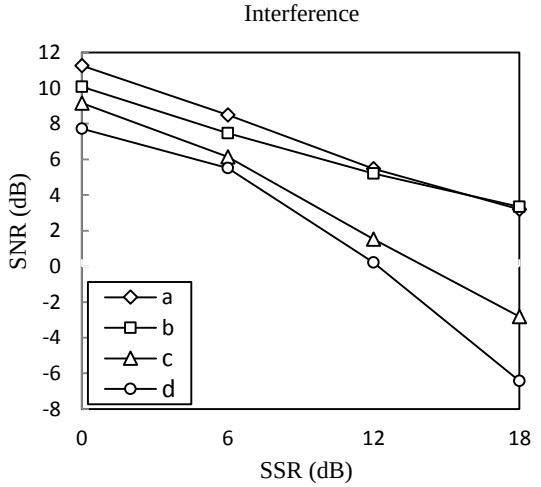

Fig. 4. averaged SNR versus SSR for separated interference speech files obtained from our proposed method (a), gain adapted MMSE estimator (b), our proposed method without gain estimation (c) and MMSE estimator (d)

VI. Conclusions

Gain difference between speakers causes improper performance of model-based single channel speech separation methods. In this paper we proposed a new VQ-based method to compensate this difference. In our proposed method, separation process is performed at the sub-section levels, gains of the speakers are estimated in the first sub-section and the estimated gains are used to estimate the feature vectors of the speakers in each sub-section. Experimental results show that our proposed method outperforms the gain adapted MMSE estimator presented in [15].

References

[1] G. Hu, D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation", IEEE Transactions on Neural Networks, vol. 15, no. 5, pp. 1135-1150, 2004.

[2] D. L. Wang, G. J. Brown, Computational Auditory Scene Analysis: Principles, Algorithms, and Applications, Wiley-IEEE Press, 2006.

[3] L. Y. Gu, R .M. Stern, "Single-Channel Speech Separation Based on Modulation Frequency", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 25-28, 2008.

[4] M. H. Radfar, R. M. Dansereau and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," EURASIP Journal on Audio, Speech and Music Processing, pp.1-15, 2007.

[5] M. Wu, D. L. Wang and G. J. Brown, "A multipitch tracking algorithm for noisy speech," IEEE Transactions on Speech and Audio Processing, vol. 11, no. 3, pp. 229-24, 2003.

[6] M .G. Christensen and A. Jakobsson, "Multi-Pitch Estimation," Synthesis Lectures on Speech and Audio Processing. Morgan and Claypool Publishers, San Rafael, CA, USA, pp. 1-24, 2009.

[7] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," IEEE Transactions on Speech and Audio Processing, vol. 8, no. 6, pp. 708-716, 2000.

[8] S. Srinivasan and D. Wang, 2008. "A model for multitalker speech perception," Journal of Acoustical Society of America, vol. 124, no. 5, pp. 3213-3224.

[9] P. Mowlaee, A. Sayadiyan and H. Sheikhzadeh, "Evaluating single channel separation performance in transform domain," Journal of Zhejiang University Science-C, Engineering Springer-Verlag, vol. 11, no. 3, pp. 160–174, March. 2010.

[10] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "Speaker independent model based single channel speech separation," Neurocomputing, vol. 72, no. 1-3, pp. 71-78, Dec. 2008.

[11] A. M. Reddy and B. Raj, "Soft mask methods for single channel speaker separation," IEEE Transactions on Audio, Speech and Language Processing., vol. 15, no. 6, pp. 1766-1776, 2007.

[12] M. H. Radfar and R. M. Dansereau, "Single channel speech separation using soft mask filtering", IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.

[13] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "A novel low complexity VQ-based single channel speech separation technique," in Proc. IEEE ISSPT06, Aug. 2006.

[14] M. J. Reyes-Gomez, D. Ellis, and N. Jojic, "Multiband audio modeling for single channel acoustic source separation," in Proc. ICASSP'04, vol. 5, pp. 641–644, May. 2004.

[15] M. H. Radfar, R. M. Dansereau, W.-Y. Chan, "Monaural speech separation based on gain adapted minimum mean square error estimation", Journal of Signal Processing Systems, vol. 61, no. 1, pp. 21-37, 2010.

[16] M. H. Radfar, A. H. Banihashemi, R. M. Dansereau, A. Sayadiyan, "A non-linear minimum mean square error estimator for the mixture- maximization approximation", Electronic Letters, vol. 42, no. 12, pp. 75–76, 2006.

[17] A. Gersho, R. M. Gray, Vector Quantization and Signal Compression. Norwell, MA: Kluwer, 1992.

[18] S. Boyd, L. Vandenberghe, Convex Optimization. Cambridge University Press, 2004

[19] B. Bradie, A Friendly Introduction to Numerical Analysis. Englewood Cliffs: Pearson Prentice Hall, 2006.

[20] M. P. Cooke, J. Barker, S. P. Cunningham, X. Shao, "An audiovisual corpus for speech perception and automatic speech recognition", Journal of Acoustical Society of America, vol. 120, pp. 2421–2424, 2006.

Authors' Profiles

Sonay Kammi received the B.Sc. and M.Sc. degrees in electronics engineering from Babol University of Technology in Iran. She is currently a Ph.D. student at Babol University of Technology. Her research interests include speech and image processing, mainly speech enhancement, nonlinear speech processing, machine learning applied to audio, and statistical signal modeling.

Mohammad Reza Karami received the Bs.c in electrical and electronic engineering in 1992, Ms.c of signal processing in 1994, and PhD in 1998 in biomedical engineering from I.N.P.L d'Nancy of France. He is now the Associate professor with the Department of Electrical and Computer Engineering, Babol University of Technology. His research interests include speech, image and signal processing.

How to cite this paper: Sonay Kammi, Mohammad Reza Karami,"Single Channel Speech Separation Using an Efficient Model-based Method", International Journal of Information Technology and Computer Science(IJITCS), vol.7, no.3, pp.42-47, 2015. DOI: 10.5815/ijitcs.2015.03.06

