A Bayesian Approach for Single Channel Speech Separation

Sonay Kammi and Mohammad Reza Karami

***Abstract*—This paper addresses the problem of single channel speech separation to extract and enhance the desired speech signals from mixed speech signals. We propose a new speech separation algorithm by utilizing Bayesian approach for the case in which, underlying sources are mixed at different levels of energies. This situation is not considered in many single channel speech separation methods. To validate the effectiveness of our proposed method, it is compared with a state-of-the art method which is a gain adapted Maximum Likelihood estimator. Through the experiments, we show that our proposed method outperforms the compared method.**

***Index Terms*—Single channel speech separation, bayesian approach, gain estimation.**

I. Introduction

In a natural world, a speech signal is frequently accompanied by other sound sources upon reaching auditory systems, yet listeners are capable of holding conversations in a wide range of conditions. This phenomenon is well known as the "cocktail party" effect [1]. It is valuable to make a computer have the ability of a human being to segregate the object source from other interfering sources. An effective separation system can greatly facilitate many applications, including automatic speech recognition (ASR), speaker identification, audio retrieval, digital content management, etc. Therefore, the research on speech separation gradually catches the researchers' attentions, and it has become an increasingly popular topic in the field of signal processing.

There are some general methods which are used to segregate speech, such as blind source separation [2] and spatial filtering [3]. These methods need multiple sensors. However, in many applications, e.g., telecommunication and audio retrieval, there is only one sensor, and a single channel solution is expected. Since in single channel separation cases only one sensor signal could be used, it is much harder and still an open problem for researchers to explore.

The aim of single channel speech separation (SCSS) is to divide an instantaneous linear mixture $z(t) = x(t) + y(t)$ of two signals into its underlying source signals $x(t)$ and $y(t)$. This is in general an ill-posed problem and can not be solved without further knowledge about the sources or their interrelationship. Possible SCSS methods are mainly divided into two categories: source driven [4-7] and model-based methods [12-18].

Manuscript received November 23, 2011, revised December 20, 2011.

Sonay Kammi and Mohammad Reza Karami are with the faculty of Faculty of Electrical and Computer Engineering, Babol University of Technology, Babol, Iran (e-mail: sonaykammi@yahoo.com; mkarami@nit.ac.ir).

As a major example for the first group, computational auditory scene analysis (CASA) has widely been studied [4]. Generally speaking, CASA-based methods aim at segregating audio sources based on possible intrinsic perceptual acoustic cues from speech signals [5]. For CASA systems, a reliable multi-pitch tracking component is critical to find pitch trajectories of two interfering speech signals [8]. 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 [5]: (1) Most existing pitch estimation methods perform reliably only with clean speech signals that have a single pitch track or harmonically related sinusoids [9] with almost no background interference [10]. (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 [6]. A high similarity between the interference and target pitch trajectories results in performance degradation of CASA methods [11]. (3) Because of energetic masking defined in [11], 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 [7].

As an alternative form of SCSS, the model-based SCSS is commonly employed to deliver an appropriate separation quality. These approaches incorporate prior knowledge such that the individual source characteristics are learned in a generative manner during a training phase. This speaker dependent model is used as source prior knowledge and is applied for separation without considering the interfering component. The most prominent models are vector quantization (VQ) [12], [16], Gaussian mixture models (GMM) [14], [15] and Hidden Marcov models (HMM) [17]. In most recent proposed model-based SCSS 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 this paper we propose a new model-based single channel speech separation method in which each source is modeled using a Gaussian mixture model. In this method, Bayesian

approach is used to estimate the gains of the speakers and the underlying sources.

The remaining paper is organized as follow: In section II, preliminary definitions are described. In section III, we model each source using a set of Gaussian subsources. We also use the MIXMAX approximation [19], to obtain the conditional probability density function (PDF) of the mixture when the sources are assumed to be known. In section IV, we use Bayesian approach to estimate the sources. Experimental results are reported in section V where separated speech signals obtained from proposed method are compared with those obtained from method presented in [18]. Finally, conclusions are drawn in section VI.

II. Preliminary Definitions

In this section, we express the gains of the sources in terms of the energy of the observation and the signal-to-signal ratio. Suppose that the observation signal is related to the two sources by

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

where $\{g_x, g_y\} > 0$ represents the associated source gains and it is assumed that the signals have equal power before gain scaling, $G_0^2 = \frac{1}{T}\sum_t x^2(t) = \frac{1}{T}\sum_t y^2(t)$. G_0^2 is a positive but otherwise arbitrary power parameter attributed to the source signals before they are scaled and mixed to produce the observed signal. From (1), we obtain

$$g_z^2 = \frac{1}{T}\sum_t z^2(t) = G_0^2\left(g_x^2 + g_y^2\right) + 2g_x g_y \frac{1}{T}\sum_t x(t)\,y(t) \tag{2}$$

for the power of the observed signal. Since $x(t)$ and $y(t)$ are two independent zero mean processes and since T is usually large, the second term in Eq. 2 vanishes. Thus, we obtain

$$g_z^2 \approx G_0^2\left(g_x^2 + g_y^2\right) \tag{3}$$

The signal-to-signal ratio (SSR) between the two sources given (1) is defined by

$$\text{SSR} = 10\log_{10}\frac{g_x^2}{g_y^2} \tag{4}$$

To show the accuracy of (3), plot of g_z^2 and $G_0^2(g_x^2 + g_y^2)$ versus SSR is shown in Fig. 1. From (3) and (4), we have

$$g_x \approx \frac{g_z}{G_0\sqrt{1+10^{\frac{-\text{SSR}}{10}}}} \quad \text{and} \quad g_y \approx \frac{g_z}{G_0\sqrt{1+10^{\frac{\text{SSR}}{10}}}} \tag{5}$$

Thus, g_x and g_y are determined if we know the SSR and G_0. We define a_x and a_y as below

$$a_x = \log_{10} g_x \quad \text{and} \quad a_y = \log_{10} g_y \tag{6}$$

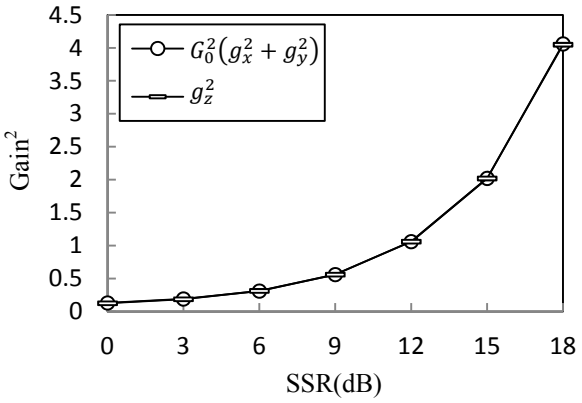

Fig. 1. Plot of g_z^2 versus $G_0^2(g_x^2 + g_y^2)$ versus SSR (dB) averaged over 10 mixed utterances.

We use (6) in the next section to obtain the relation between the log spectra of the sources and that of the observation signal.

III. Source and Mixture Modeling

A. Feature Selection

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.

$$X\left(e^{\frac{j2\pi d}{L}}\right) = \sum_{l=0}^{L-1} X(l)e^{-\frac{j2\pi ld}{L}} \qquad d = 0,1,\dots,D-1 \tag{7}$$

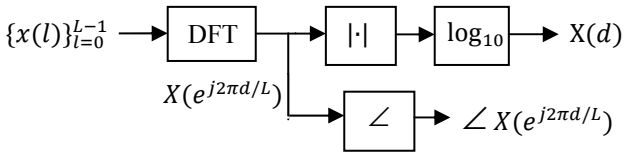

Fig. 2. Feature extraction

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

$$X(d) = \log_{10}\left|X(e^{j2\pi d/L})\right| \qquad d = 0,1,\dots,D-1 \tag{8}$$

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. 2.

B. Source Modeling

Let x^r and y^r be the N-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

$$X^r = \log_{10}(|F_D(x^r)|) \tag{9}$$

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

$$Z^r = \log_{10}(|F_D(z^r)|) \quad (11)$$

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.

In order to model the feature space (log spectral vectors) of each speaker, we use the composite source modeling (CSM) technique in which each source is represented by a finite set of statistically independent Gaussian subsources. The model is fully determined by the a priori probability of subsources and the probability of the observation occurring given the subsource. Let $S_x = \{s_x = i | i = 1,2, \dots, I\}$ and $S_y = \{s_y = j | j = 1,2, \dots J\}$ denote the set of subsources of speaker one and two, respectively. Also, let $p_{s_x}(s_x = i)$, with the constraint $\sum_i p_{s_x}(s_x = i) = 1$, and $p_{s_y}(s_y = j)$, with the constraint $\sum_j p_{s_y}(s_y = j) = 1$ denote the prior probability of subsources $s_x = i$ and $s_y = j$, respectively. For each speaker, the subsources are modeled using Gaussian distributions and its PDF is obtained in the form

$$p_{X^r|s_x}(X^r|s_x = i) = \prod_d p_{X^r|s_x}(X_d^r|s_x = i) \quad (12)$$

where

$$p_{X^r|s_x}(X_d^r|s_x = i) = \prod_d \frac{1}{\sqrt{2\pi\sigma_{x,d}^{2i}}} exp\left(-\frac{\left(X_d^r - \mu_{x,d}^i\right)^2}{2\sigma_{x,d}^{2i}}\right) \quad (13)$$

and for speaker two

$$p_{Y^r|s_y}(Y^r|s_y = j) = \prod_d p_{Y^r|s_y}(Y_d^r|s_y = j) \quad (14)$$

where

$$p_{Y^r|s_y}(Y_d^r|s_y = j) = \prod_d \frac{1}{\sqrt{2\pi\sigma_{y,d}^{2j}}} exp\left(-\frac{\left(Y_d^r - \mu_{y,d}^j\right)^2}{2\sigma_{y,d}^{2j}}\right) \quad (15)$$

where $\mu_{x,d}^i$ is the dth component of the mean vector, and $\sigma_{x,d}^{2i}$ is the dth component on the diagonal of the covariance matrix of the ith subsource for speaker one. Likewise, $\mu_{y,d}^j$ is the dth component of the mean vector, and $\sigma_{y,d}^{2j}$ is the dth component on the diagonal of the covariance matrix of the jth subsource for speaker two.

C. Mixture Modeling

To proceed further, we need to obtain the relation between the log spectra of the observation Z_d^r and those of the sources X_d^r and Y_d^r. Performing a procedure similar to the one presented in [19], this relation is given by

$$Z_d^r = max\left(X_d^r + a_x, Y_d^r + a_y\right) + e_d \quad (16)$$

in which $\max(\cdot,\cdot)$ returns the larger element and e_d is an approximation error which is modeled by a zero mean white Gaussian noise with variance σ_d^2 in the form

$$p_e(e_d) = \frac{1}{\sqrt{2\pi\sigma_d^2}} \exp\left(\frac{-e_d^2}{2\sigma_d^2}\right) \quad (17)$$

From (16) and (17), we obtain

$$p_{Z^r|X^r,Y^r,\mathrm{SSR}}(Z_d^r|X_d^r, Y_d^r, \mathrm{SSR}) = p_e(e_d) \quad (18)$$

where

$$p_e(e_d) = \frac{1}{\sqrt{2\pi\sigma_d^2}} \times exp\left(\frac{\left(Z_d^r - max\left(X_d^r + a_x, Y_d^r + a_y\right)\right)^2}{2\sigma_d^2}\right) \quad (19)$$

Note that for notational simplicity, we hereafter drop the subscript which appears in the definition of the PDF where, for instance, we denote $p_a(a)$ by $p(a)$unless otherwise noted. The obtained equations are used in the next section to estimate the sources.

IV. Derivations of the Estimator Using Bayesian Approach

A. Problem Formulation

The main goal of this paper is to obtain an estimate of $X^r(d)$ and $Y^r(d)$, that is $\hat{X}^r(d)$ and $\hat{Y}^r(d)$, given $Z^r(d)$ as well as the probability density function (PDF) of $X^r(d)$ and $Y^r(d)$. To do this, we propose a method based on the Bayesian approach. In this framework, the estimate of $X^r(d)$ is given by

$$\hat{X}_d^r = arg \max_{X_d^r} p(Z_d^r|X_d^r, Y_d^r) p(X_d^r) p(Y_d^r) \quad (20)$$

B. Derivations of the Estimator Given Subsources and SSR

In order to derive the estimator, we assume that the sources X^r and Y^r are in subsources i^* and j^*, respectively. Also, we assume signal-to-signal ratio SSR* is given which form (5) and (6) means that a_x^* and a_y^* are known. Thus, the Bayesian estimator to estimate $X^r$is given by

$$\hat{X}_d^r = arg \max_{X_d^r} p(Z_d^r|X_d^r, Y_d^r, \mathrm{SSR}^*) p(X_d^r|s_x = i^*) p\left(Y_d^r|s_y = j^*\right) \quad (21)$$

We take the log of (21) because working in log-space is easier and does not change the location of highest probability

$$\hat{X}_d^r = arg\max_{X_d^r}(\log p(Z_d^r|X_d^r, Y_d^r, \text{SSR}^*) + \log p(X_d^r|s_x = i^*) + \log p\left(Y_d^r\middle|s_y = j^*\right)) \quad (22)$$

Replacing $p(Z_d^r|X_d^r, Y_d^r, \text{SSR}^*)$ with (18), $p(X_d^r|s_x = i^*)$ with (13) and $p\left(Y_d^r\middle|s_y = j^*\right)$ with (15) and dropping constants including, $\log \sigma_{x,d}^{i^*}$, $\log \sigma_{y,d}^{j^*}$, $\log \sigma_d$ and $3/2 \log 2\pi$ (because they are independent of X_d^r and Y_d^r) we obtain:

$$\hat{X}_d^r = arg\min_{X_d^r}\left(\frac{\left(Z_d^r - max\left(X_d^r + a_x^*, Y_d^r + a_y^*\right)\right)^2}{2\sigma_d^2} + \frac{\left(X_d^r - \mu_{x,d}^{i^*}\right)^2}{2\sigma_{x,d}^{2i^*}} + \frac{\left(Y_d^r - \mu_{y,d}^{j^*}\right)^2}{2\sigma_{y,d}^{2j^*}}\right) \quad (23)$$

It should be noted that minimization obtained from the fact that maximizing an exponential function with a negative argument corresponds to minimizing the argument of the exponential function. To find the minimum of (16) we differentiate (16) with respect to X_d^r and set the result to zero:

$$\frac{\partial}{\partial X_d^r}\left(\frac{\left(Z_d^r - max\left(X_d^r + a_x^*, Y_d^r + a_y^*\right)\right)^2}{2\sigma_d^2} + \frac{\left(X_d^r - \mu_{x,d}^{i^*}\right)^2}{2\sigma_{x,d}^{2i^*}} + \frac{\left(Y_d^r - \mu_{y,d}^{j^*}\right)^2}{2\sigma_{y,d}^{2j^*}}\right) = 0 \quad (24)$$

To perform the derivation we make this assumption that if $\mu_{x,d}^{i^*} + a_x^* > \mu_{y,d}^{j^*} + a_y^*$ then $max\left(X_d^r + a_x^*, Y_d^r + a_y^*\right) = X_d^r$ and vice versa. Thus performing the derivation introduced in (22), we obtain

$$\hat{X}_d^r = \left[W\left(\sigma_{x,d}^{2i^*}, \sigma_d^2\right)(Z_d^r - a_x^*) + W\left(\sigma_d^2, \sigma_{x,d}^{2i^*}\right)\mu_{x,d}^{i^*}\right]\Omega\left(\left(\mu_{x,d}^{i^*} + a_x^*\right) - \left(\mu_{y,d}^{j^*} + a_y^*\right)\right) + \left[\mu_{x,d}^{i^*}\right]\Omega\left(\left(\mu_{y,d}^{j^*} + a_y^*\right) - \left(\mu_{x,d}^{i^*} + a_x^*\right)\right) \quad (25)$$

where $\mathrm{W(a,b)} = \frac{\mathrm{a}}{\mathrm{a+b}}$ and also, $\Omega(\mathrm{a} - \mathrm{b})$ denotes the step function in which if $\mathrm{a} \geq \mathrm{b}$ then $\Omega(\mathrm{a} - \mathrm{b}) = 1$ and if $\mathrm{a} < b$ then $\Omega(\mathrm{a} - \mathrm{b}) = 0$.

Likewise, when $\mu_{\mathrm{y,d}}^{\mathrm{j}^*} + a_y^* > \mu_{\mathrm{x,d}}^{\mathrm{i}^*} + a_x^*$, we have $max\left(X_d^r + a_x^*, Y_d^r + a_y^*\right) = Y_d^r$. To obtain Bayesian estimation of Y_d^r, we minimize (23) with respect to Y_d^r. Taking the partial derivative of (23) with respect to Y_d^r and setting to zero yields

$$\hat{Y}_d^r = \left[W\left(\sigma_{y,d}^{2j^*}, \sigma_d^2\right)\left(Z_d^r - a_y^*\right) + W\left(\sigma_d^2, \sigma_{y,d}^{2j^*}\right)\mu_{y,d}^{j^*}\right]\Omega\left(\left(\mu_{y,d}^{j^*} + a_y^*\right) - \left(\mu_{x,d}^{i^*} + a_x^*\right)\right) + \left[\mu_{y,d}^{j^*}\right]\Omega\left(\left(\mu_{x,d}^{i^*} + a_x^*\right) - \left(\mu_{y,d}^{j^*} + a_y^*\right)\right) \quad (26)$$

In this way, we obtain the Bayesian approach based estimates of X_d^r and Y_d^r in terms of $\{\mu_{x,d}^{i^*}, \mu_{y,d}^{j^*}, \sigma_{x,d}^{2i^*}, \sigma_{y,d}^{2j^*}, \sigma_d^2, a_x^*, a_y^*\}$ in which, $\{\mu_{x,d}^{i^*}, \mu_{y,d}^{j^*}, \sigma_{x,d}^{2i^*}, \sigma_{y,d}^{2j^*}, \sigma_d^2\}$ are available as a priori knowledge of the sources and the noise model and $\{a_x^*, a_y^*\}$ are obtained fromSSR^*.

C. Estimating SSR and Detecting the Subsources

The Bayesian approach based estimators (25) and (26) were obtained with the assumption that signal-to-signal ratio SSR^* and subsources $s_x = i^*$ and $s_y = j^*$ are known in advance. In this section, the two subsources $s_x = i^*$ and $s_y = j^*$ along with a value of the SSR, SSR^*, which maximize the PDF of the subsources and SSR given observation over all frames are estimated. Mathematically, this procedure which is Bayesian estimation is expressed by

$$\{i^*, j^*, \text{SSR}^*\} = arg\max_{\text{SSR}}\sum_{r=1}^{R} arg\max_{i,j}\log p\left(s_x = i, s_y = j, \text{SSR}\middle|Z^r\right) \quad (27)$$

where R denotes the number of frames. Using the Bayes' rule we have

$$\{i^*, j^*, \text{SSR}^*\} = arg\max_{\text{SSR}}\sum_{r=1}^{R} arg\max_{i,j}(\log p\left(Z^r\middle|s_x = i, s_y = j, \text{SSR}\right) + \log p(s_x = i) + \log p\left(s_y = j\right)) \quad (28)$$

where $p(s_x = i)$ and $p\left(s_y = j\right)$ are the prior probability of subsources and $p\left(Z^r\middle|s_x = i, s_y = j, \text{SSR}\right)$ represents the PDF of Z^r given the subsources $s_x = i$ and $s_y = j$ and SSR. Performing a procedure similar to the one presented in [15], $p\left(Z^r\middle|s_x = i, s_y = j, \text{SSR}\right)$ can be approximated in the following form

$$p\left(Z^r\middle|s_x = i, s_y = j, \text{SSR}\right) \approx \prod_d \frac{1}{\sqrt{2\pi\sigma_{d\,max}^2}} \times exp\left(-\frac{\left(Z_d^r - max\left(\mu_{x,d}^i + a_x, \mu_{y,d}^j + a_y\right)\right)^2}{2\sigma_{d\,max}^2}\right) \quad (29)$$

where $\sigma_{d\,max}^2$ is the varians of the mixture (i.e. $s_x = i$ or $s_y = j$) whose mean is greater than the other.

The procedure performed in (27) can be explained as follows. Let $\mathrm{S} = \{\mathrm{SSR}_1, \ldots, \mathrm{SSR_m}, \ldots, \mathrm{SSR_M}\}$ be a set of possible SSRs. We first set the SSR to one of these values and then for each frame estimate the subsouces which maximize $\log p\left(s_x = i, s_y = j, \text{SSR}\middle|Z^r\right)$ for that frame. The SSR which maximizes $\log p\left(s_x = i, s_y = j, \text{SSR}\middle|Z^r\right)$ for all frames along with decoded subsources corresponding to this SSR are selected and used to recover the sources.

D. Synthesizing Estimated Speech 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

$$\hat{x} = F_D^{-1}\left(10^{\hat{X}}\exp\left[(-1)^{\frac{1}{2}}\angle F_D(z)\right]\right) \quad (30)$$

and

$$\hat{y} = F_D^{-1}\left(10^{\hat{Y}}\exp\left[(-1)^{\frac{1}{2}}\angle F_D(z)\right]\right) \quad (31)$$

where $\angle$ denotes the phase operator, exp[·] 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. Experiments and Results

In this section, we perform experiments to validate the effectiveness of the proposed method and also compare it with the technique presented in [18] in which ML estimation is used to estimate the gains of the sources and then, a binary mask filter is used to estimate the sources. Speech files considered for the experiments are selected from the database presented in [20]. The database consists of speech files of 34 speakers. For the test phase, 10 randomly selected pair of speech files of speakers that were not included in the training data set are selected and mixed at signal-to-signal ratios equal to 0, 6, 12, 18 dB. One of the speech signals in each mixture is treated as the target signal while the other one as the interference signal. Throughout the experiments, a hamming window of length 50 ms with frame shift equal to 20 ms in the training phase and 10 ms in the test phase is used to segment the speech files. Also a Hann window is used in the overlap-add method for synthesizing the separated speech signals. The sampling rate is decreased to 8 kHz from the original 25 kHz. In order to generate the CSMs for log spectral vectors, the training data vectors are transformed to the log spectral domain using a 512-point discrete Fourier transform, such that after discarding the symmetric portion, the log spectral vectors with a dimension of 257 are obtained. Afterwards, the training data set for speaker is clustered into 128 subsources using Linde-Buzo-Gray (LBG) algorithm [21].

In order to evaluate the proposed technique, the signal-to-noise ratio (SNR) between the estimated and original signals in the time domain is used. The SNR for the estimated signal of the speaker x is defined as

$$\mathrm{SNR}_x = 10\log_{10}\left[\frac{\sum_t (x(t))^2}{\sum_t (x(t)-\hat{x}(t))^2}\right] \quad (32)$$

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

Experimental results are shown in Fig. 3 and Fig. 4. In these figures, the SNR versus SSR averaged over 10 separated target and interference speech files are shown for our proposed method and the technique given in [18].

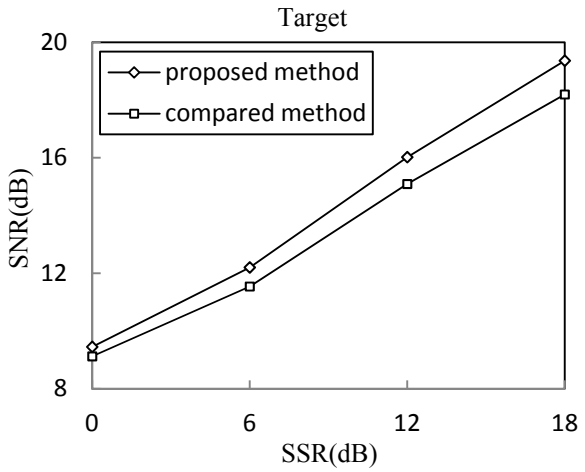

Fig. 3. SNR versus SSR averaged over 10 separated target speech files obtained from proposed method and the method presented in [18]

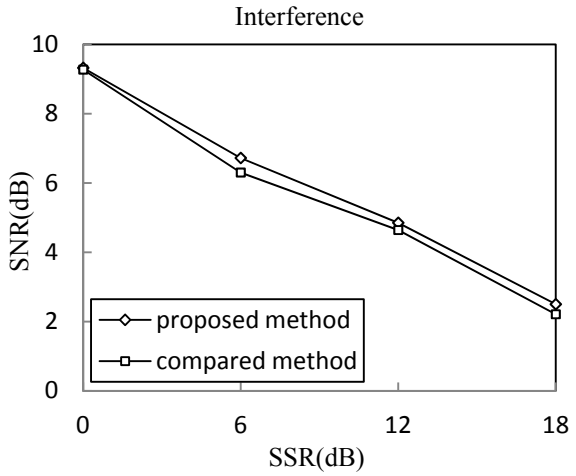

Fig. 4. SNR versus SSR averaged over 10 separated interference speech files obtained from proposed method and the method presented in [18]

As it is obvious from the figures, our proposed method outperforms the method given in [18].

VI. Conclusion

In this paper, a Bayesian approach has been proposed for single channel speech separation. In this approach, not only sources, but also associated gains are estimated. Thus, the proposed technique can be applied for situations in which training and test data are recorded under different conditions. Through the experiments, we have shown that our proposed technique outperforms the technique presented in [18].

References

[1] C. Cherry, "Some experiments in the recognition of speech with one and two ears," *J. Acoust. Soc. Amer.*, vol. 25, pp. 975–981, 1953.

[2] A. K. Barros, T. Rutkowski, F. Itakura, and N. Ohnishi, "Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets," *IEEE Trans. Neural Netw.*, vol. 13, no. 4, pp. 888–893, Jul. 2002.

[3] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

[4] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neur. Networks*, vol. 15, no. 5, pp. 1135-1150, 2004.
[5] D. Wang and G. J. Brown, Eds., Computational Auditory Scene Analysis: Principles, Algorithms, and Applications, ser. IEEE Press. J. Wiley and Sons Ltd, 2006.
[6] L. Y. Gu and R .M. Stern, "Single-Channel Speech Separation Based on Modulation Frequency," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 25-28, 2008.
[7] 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 J. Audio Speech Music Process,* pp.1-15, 2007.
[8] M. Wu, D. L. Wang and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 229-24, 2003.
[9] 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.
[10] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 708-716, 2000.
[11] S. Srinivasan and D. Wang, 2008. "A model for multitalker speech perception," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 3213-3224.
[12] 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.
[13] 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.
[14] A. M. Reddy and B. Raj, "Soft mask methods for single channel speaker separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 6, pp. 1766-1776, 2007.
[15] 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.
[16] 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.
[17] 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.
[18] M. H. Radfar and R. M. Dansereau, "Long-term gain estimation in model-based single channel speech separation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA2007), New Paltz, New York,* October. 2007.
[19] M. H. Radfar, A. H. Banihashemi, R. M. Dansereau, and A. Sayadiyan, "A non-linear minimum mean square error estimator for the mixture-maximization approximation," *Electron. Lett.*, vol. 42, no. 12, pp. 75–76, Jun. 2006.
[20] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audiovisual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 120, pp. 2421–2424, Nov. 2005.
[21] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer, 1992.