Single Channel Speech Separation Using Maximum a Posteriori Estimation

M. H. Radfar and R. M. Dansereau

Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada
{radfar,rdanse}@sce.carleton.ca

10.21437/Interspeech.2007-342

Abstract

We present a new approach for separating two speech signals when only a single recording of their additive mixture is available. In this approach, log spectra of the sources are estimated using maximum a posteriori estimation given the mixture's log spectrum and the probability density functions of the sources. It is shown that the estimation leads to a two-state, non-linear filter whose states are controlled by the means of the sources. The first state of the filter is expressed using a combination of two Wiener filters whose parameters are controlled by the means and variances of the sources and noise variance and the second state is expressed by the means of the sources. Through the experiments, conducted on a wide variety of mixtures, we show that the MAP based estimator outperforms the methods which use binary mask filtering or Wiener filtering for the separation task.

Index Terms:source separation, single channel speech separation, maximum a posteriori estimation, binary mask.

1. Introduction

Model-based single channel speech separation (SCSS) [1] 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. In these techniques, patterns of the sources – one from each source– whose combination best match the mixture are selected and used to recover the sources. Two important aspects of model-based SCSS are: (i) the applied modeling technique and (ii) the estimation technique. As far as the modeling techniques are concerned many approaches have been proposed mainly based on: vector quantization [2, 3], Gaussian mixture modeling [4], hidden Markov models [5, 6], non-negative matrix factorization [7], and independent component analysis [8].

Given the individual speakers' models, an estimation technique is applied which is usually aim at optimizing a target function based on the maximum likelihood (ML) [8], minimum mean square error maximum (MMSE) [9], or maximum a posteriori (MAP) [4] approaches. The result of this optimization often leads either to the direct estimate of the sources [7–9] or to filters which when imposed on the mixture obtains an estimate of the desired source [2–6]. In current model-based SCSS, the Wiener filter or the binary mask filter are frequently used. The Wiener filter, a well-studied tool in estimation theory, has been used in either time [4] or frequency [6] based separation techniques. On the other hand, the binary mask filter [2, 3, 5] is only used in the log frequency domain where it is assumed that at each frequency band the log spectral component with higher amplitude masks the other component.

The authors would like to thank the Natural Sciences and Engineering Research Council (NSERC) of Canada which funded this research.

In this paper, we propose a new filtering approach which is the result of the MAP based estimate of the log spectra of the sources given the mixture and with the assumption that the sources and mixture are related through the use of mixture-maximization (MIXMAX) approximation [10]. In this technique, each source is modeled using a simplified hidden Markov model in which the dynamic process between the states are not considered for reducing computational complexity.Then, given the states of each of HMMs associated with the sources, the MAP estimator is derived which leads to a non-linear filter which is expressed using a combination of two Wiener filters whose parameters are controlled by the mean and variance of states and noise variance. Finally, two states which maximize the log likelihood of the mixture PDF are decoded and their parameters are passed to the filters.

The rest of this paper is organized as follows. In Sec.2, preliminary definitions and source-observation modleing are described. The source estimation approach is presented in Sec.3. In Sec.4, the state estimation is introduced and then experimental set-up, a short description of the compared methods and the experimental results are given in Sec. 5. Finally, conclusion are drawn are Sec.6.

2. Preliminary definitions and models

2.1. Problem Formulation

Let $\mathbf{s}_k^t = [s_k^t(0), \ldots, s_k^t(n), \ldots, s_k^t(N-1)]^\top$, where $k \in \{1, 2\}$ and $\top$ denotes the transpose operator, be a frame of length N of the speech signal of the k^{th} speaker in the time domain. Let $\mathbf{z}^t = \mathbf{s}_1^t + \mathbf{s}_2^t$ be the corresponding frame of the observation signal. Let $\mathbf{s}_k = [s_k(0), \ldots, s_k(d), \ldots, s_k(D-1)]^\top$ and $\mathbf{z} = [z(0), \ldots, z(d), \ldots, z(D-1)]^\top$ be the log spectral vectors obtained by taking the logarithm from the magnitude of the D-point discrete Fourier transform of $\mathbf{s}_k^t$ and $\mathbf{z}$, respectively. More specifically, $\mathbf{s}_k = \log_{10}|\mathcal{F}_D(\mathbf{s}_k^t)|$ $\mathbf{z} = \log_{10}|\mathcal{F}_D(\mathbf{z}^t)|$ where $\mathcal{F}_D(\cdot)$ denotes the D-point discrete Fourier transform, $|\cdot|$ denotes the magnitude operator, and $\log_{10}$ denotes the base 10 logarithm operator.

The main goal of this paper is to obtain an estimate of $\mathbf{s}_1$ and $\mathbf{s}_2$, that is $\hat{\mathbf{s}}_1$ and $\hat{\mathbf{s}}_2$, given $\mathbf{z}$ as well as the probability density function (PDF) of $\mathbf{s}_1$ and $\mathbf{s}_2$. To do this, we propose a method based on the maximum a posteriori (MAP) approach. In this framework, the estimate of $\mathbf{s}_1$ is given by[1]

$$\hat{\mathbf{s}}_1 = \arg\max_{\mathbf{s}_1} \ln p_{\mathbf{z}|\mathbf{s}_1,\mathbf{s}_2}(\mathbf{z}|\mathbf{s}_1,\mathbf{s}_2) p_{\mathbf{s}_1}(\mathbf{s}_1) p_{\mathbf{s}_2}(\mathbf{s}_2) \qquad (1)$$

where $\ln = \log_e$ and $p_{\mathbf{z}|\mathbf{s}_1,\mathbf{s}_2}(\mathbf{z}|\mathbf{s}_1,\mathbf{s}_2)$ is the conditional PDF of $\mathbf{z}$ given the sources $\mathbf{s}_1$ and $\mathbf{s}_2$. Also $p_{\mathbf{s}_1}(\mathbf{s}_1)$ and $p_{\mathbf{s}_2}(\mathbf{s}_2)$ represent the PDFs of the sources. From (1), it is readily seen that we need

[1]Note that the estimate of $\mathbf{s}_2$ is obtained in a similar fashion only with swapping $\mathbf{s}_1$ with $\mathbf{s}_2$.

to model $p_{\mathbf{z}|\mathbf{s}_1,\mathbf{s}_2}(\mathbf{z}|\mathbf{s}_1,\mathbf{s}_2)$, $p_{\mathbf{s}_1}(\mathbf{s}_1)$, and $p_{\mathbf{s}_2}(\mathbf{s}_2)$, a topic which is discussed in Sec. 2.2.

2.2. Modeling the PDFs of the sources and observation

We model each source using a hidden Markov model (HMM) whose transition matrix is assumed to consist of equal probable elements. Therefore, the source $\mathbf{s}_1$ is modeled using an HMM with the state space $\mathcal{X}_1 = \{x_1^i, i = 1, \ldots, I\}$. In each state, the source $\mathbf{s}_1$ is assumed to be modeled with a Gaussian PDF with a diagonal covariance matrix, that is

$$p_{\mathbf{s}_1|x_1^i}(\mathbf{s}_1|x_1^i) = \prod_{d=1}^{D} p_{s_1|x_1^i}\big(s_1(d)|x_1^i\big) \tag{2}$$

where

$$p_{s_1|x_1^i}\big(s_1(d)|x_1^i\big) = \frac{1}{\sigma_1^i(d)\sqrt{2\pi}} \exp\frac{-1}{2}\Big(\frac{s_1(d)-\mu_1^i(d)}{\sigma_1^i(d)}\Big)^2 \tag{3}$$

in which $\mu_1^i(d)$ is the d^{th} component of the mean vector and $\sigma_1^i(d)$ is the d^{th} component on the diagonal of the covariance matrix of the i^{th} state. Likewise, the source $\mathbf{x}_2$ is modeled with an HMM with state space $\mathcal{X}_2 = \{x_2^j, j = 1, \ldots, J\}$. In each state, the source $\mathbf{s}_2$ is assumed to be modeled with a Gaussian PDF with a diagonal covariance matrix, that is

$$p_{\mathbf{s}_2|x_2^j}(\mathbf{s}_2|x_2^j) = \prod_{d=1}^{D} p_{s_2|x_2^j}\big(s_2(d)|x_2^j\big) \tag{4}$$

where

$$p_{s_2|x_2^j}\big(s_2(d)|x_2^j\big) = \frac{1}{\sigma_2^j(d)\sqrt{2\pi}} \exp\frac{-1}{2}\Big(\frac{s_2(d)-\mu_2^j(d)}{\sigma_2^j(d)}\Big)^2 \tag{5}$$

in which $\mu_2^j(d)$ is the d^{th} component of the mean vector and $\sigma_2^j(d)$ is the d^{th} component on the diagonal of the covariance matrix of j^{th} state.

To proceed further, we need to obtain the relation between the log-spectra of the observation $z(d)$ and those of the sources $s_1(d)$ and $s_2(d)$. This relation can be expressed by the mixture-maximization (MIXMAX) approximation [10] in the form

$$z(d) = \max\big(s_1(d), s_2(d)\big) + e(d) \quad d = 1, \ldots, D \tag{6}$$

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

$$p_e\big(e(d)\big) = \frac{1}{\sigma(d)\sqrt{2\pi}} \exp\frac{-1}{2}\Big(\frac{e(d)}{\sigma(d)}\Big)^2 \tag{7}$$

From (6) and (7), we obtain

$$p_{\mathbf{z}|\mathbf{s}_1,\mathbf{s}_2}(\mathbf{z}|\mathbf{s}_1,\mathbf{s}_2) = \prod_{d=1}^{D} p_e\big(e(d)\big) \tag{8}$$

where

$$p_e\big(e(d)\big) = \frac{1}{\sigma(d)\sqrt{2\pi}} \exp\frac{-1}{2}\Big(\frac{z(d)-\max\big(s_1(d),s_2(d)\big)}{\sigma(d)}\Big)^2. \tag{9}$$

In the next section we use (2),(3), (4), (5), (8), and (9) to compute (1).

3. Derivation of the MAP Estimator given the states x_1^i and x_2^j

In order to derive the MAP estimator, we assume that the source $\mathbf{s}_1$ and $\mathbf{s}_2$ are in states i and j, respectively. Thus, the MAP estimator (1) to estimate $\mathbf{s}_1$ is rewritten by

$$\hat{\mathbf{s}}_1 = \arg\max_{\mathbf{s}_1} \ln p_{\mathbf{z}|\mathbf{s}_1,\mathbf{s}_2}(\mathbf{z}|\mathbf{s}_1,\mathbf{s}_2) p_{\mathbf{s}_1|x_1^i}(\mathbf{s}_1|x_1^i) p_{\mathbf{s}_2|x_2^j}(\mathbf{s}_2|x_2^j). \tag{10}$$

Substituting $p_{\mathbf{z}|\mathbf{s}_1,\mathbf{s}_2}(\mathbf{z}|\mathbf{s}_1,\mathbf{s}_2)$, $p_{\mathbf{s}_1|x_1^i}(\mathbf{s}_1|x_1^i)$, and $p_{\mathbf{s}_2|x_2^j}(\mathbf{s}_2|x_2^j)$ respectively with (8) and (4), and (2), we have

$$\begin{aligned}\hat{\mathbf{s}}_1 = \arg\max_{\mathbf{s}_1} &\sum_{d=1}^{D} \ln p_e\big(e(d)\big) \\ &+ \ln p_{s_1|x_1^i}\big(\mathbf{s}_1(d)|x_1^i\big) + \ln p_{s_2|x_2^j}\big(\mathbf{s}_2(d)|x_2^j\big).\end{aligned} \tag{11}$$

Then, substituting $p_e\big(e(d)\big)$, $p_{s_1|x_1^i}\big(s_1(d)|x_1^i\big)$, and $p_{s_2|x_2^j}\big(s_2(d)|x_2^j\big)$ respectively with (9), (5), and (3) and imposing the ln operator into the PDFs, we arrive at the following minimization problem

$$\begin{aligned}\hat{\mathbf{s}}_1 = \arg\min_{\mathbf{s}_1} &\sum_{d=1}^{D} \frac{1}{2}\Big(\frac{z(d)-\max\big(s_1(d),s_2(d)\big)}{\sigma(d)}\Big)^2 \\ &+ \frac{1}{2}\Big(\frac{s_1(d)-\mu_1^i(d)}{\sigma_1^i(d)}\Big)^2 + \frac{1}{2}\Big(\frac{s_2(d)-\mu_2^j(d)}{\sigma_2^j(d)}\Big)^2 - A(d)\end{aligned} \tag{12}$$

where $A(d) = \ln\sigma(d) + \ln\sigma_1(d) + \ln\sigma_2(d) + \frac{3}{2}\ln 2\pi$ is a term independent of $s_1(d)$ and $s_2(d)$. 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. (12) is minimized by taking the derivation of (12) with respect to $s_1(d)$, that is

$$\begin{aligned}\frac{\partial}{\partial s_1(d)}\Big[\Big(&\frac{z(d)-\max\big(s_1(d),s_2(d)\big)}{\sigma(d)}\Big)^2 \\ &+ \Big(\frac{s_1(d)-\mu_1^i(d)}{\sigma_1^i(d)}\Big)^2 + \Big(\frac{s_2(d)-\mu_2^j(d)}{\sigma_2^j(d)}\Big)^2\Big] = 0.\end{aligned} \tag{13}$$

To perform the derivation we make this assumption that if $\mu_1^i(d) \geq \mu_2^j(d)$ then $\max\big(s_1(d),s_2(d)\big) = s_1(d)$ and vice versa. Thus performing the derivation introduced in (13), we obtain

$$\begin{aligned}\hat{s}_1(d) = &\big[W\big(\sigma_1^{2i}(d),\sigma^2(d)\big)z(d) \\ &+ W\big(\sigma^2(d),\sigma_1^{2i}(d)\big)\mu_1^i(d)\big]\Omega\big(\mu_1^i(d)-\mu_2^j(d)\big) \\ &+ \big[\mu_1^i(d)\big]\Omega\big(\mu_2^j(d)-\mu_1^i(d)\big) \quad d = 1, \ldots, D\end{aligned} \tag{14}$$

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

Similarly, the estimate of $\mathbf{s}_2$ using the MAP estimator is obtained by

$$\begin{aligned}\hat{s}_2(d) = &\big[W\big(\sigma_2^{2j}(d),\sigma^2(d)\big)z(d) \\ &+ W\big(\sigma^2(d),\sigma_2^{2j}(d)\big)\mu_2^j(d)\big]\Omega\big(\mu_2^j(d)-\mu_1^i(d)\big) \\ &+ \big[\mu_2^j(d)\big]\Omega\big(\mu_1^i(d)-\mu_2^j(d)\big) \quad d = 1, \ldots, D.\end{aligned} \tag{15}$$

In this way, we obtain the MAP based estimates of $\mathbf{s}_1$ and $\mathbf{s}_2$ in terms of the means and variances of the sources' PDFs and the noise variance, i.e. $\{\mu_1^i(d), \mu_2^j(d)\sigma_1^{2i}(d), \sigma_2^{2j}(d), \sigma^2(d)\}$, all of which are available as *a priori* knowledge of the sources and the noise model.

4. Detecting the states x_1^i and x_2^j

The MAP based estimators (14) and (15) were obtained with the assumption that the states x_1^i and x_2^j are known in advance. In practice, we however do not know at which states the sources are at a given time instant. Therefore, we select two states which maximize the PDF of the observation signal. Mathematically, that means

$$\{x_1^{i^*}, x_2^{j^*}\} = \arg\max_{x_1^i, x_2^j} \ln \prod_{d=1}^{D} p_{z|x_1^i, x_2^j}\big(z(d)|x_1^i, x_2^j\big) \quad (16)$$

where $p_{z|x_1^i, x_2^j}\big(z(d)|x_1^i, x_2^j\big)$ represents the PDF of $z(d)$ given the states x_1^i and x_2^j. From (6) and with ignoring the modeling error, $p_{z|x_1^i, x_2^j}\big(z(d)|x_1^i, x_2^j\big)$ can be approximated in the following form

$$p_{z|x_1^i, x_2^j}\big(z(d)|x_1^i, x_2^j\big) \approx p_{s_1|x_1^i}\big(z(d)|x_1^i\big)\Omega\big(\mu_1^i(d) - \mu_2^j(d)\big) + p_{s_2|x_2^j}\big(z(d)|x_2^j\big)\Omega\big(\mu_2^j(d) - \mu_1^i(d)\big). \quad (17)$$

Note that $p_{s_1|x_1^i}\big(s_1(d)|x_1^i\big)$ and $p_{s_2|x_2^j}\big(s_2(d)|x_2^j\big)$ are computed in the training phase and are considered *a priori* knowledge of the sources. Hence, the two states which maximize (16) are detected and the corresponding parameters of the detected states, i.e. $\{\mu_1^{i^*}(d), \mu_2^{j^*}(d), \sigma_1^{2i^*}(d), \sigma_2^{2j^*}(d)\}$, along with the noise variance $\sigma^2(d)$ are passed to (14) and (15) to estimate $\mathbf{s}_1$ and $\mathbf{s}_2$.

5. Experiments

In this section, we perform experiments to evaluate the proposed technique and also compare it with the two state-of-the-art techniques. In Sec. 5.1, experimental set-up are explained, then in Sec. 5.2, the techniques with which our technique is compared are briefly described. Finally, the experimental results are reported in Sec. 5.3.

5.1. Experimental set-up

Six speakers, three males and three females, are chosen from the database introduced in [11]. In this database, each speaker has uttered 500 sentences. we choose 450 sentences of each speaker for the training phase in which $\{\mu_1^{i^*}(d), \mu_2^{j^*}(d), \sigma_1^{2i^*}(d), \sigma_2^{2j^*}(d)\}$ are computed. Five sentences of each speaker , not in the training data set, are chosen and mixed at the signal-to-signal ratio (SSR) equal to zero dB to form 5 female-female (I), 5 male-male (II), and 5 female-male (III) mixtures. The estimated speech files , 10 female speech files from I, 10 male speech files from II, 5 male and 5 female speech files from III, are compared with the original speech file using SNR criterion.

Throughout the experiments, a Hamming window of length 30 msec with frame rate equal to 10 msec is used to segment the speech files either in the training or test phase. Performing segmentation on the training data, about 70000 vectors are obtained for each speaker. The training data are then transformed to the spectral domain using a 512-point discrete Fourier transform such that after discarding the symmetric portion, the spectral vectors with a dimension of $D = 256$ are obtained. Afterward, the training data set for each speaker are clustered into 128 states using the Linde-Buzo-Gray (LBG) algorithm algorithm. Then the mean vector and variance of each state are computed. The prior probability of each state is calculated as the ratio between the number of data in each cluster and the number of all training data sets. As mentioned earlier, we do not consider the dynamic relation between the states, or, in other words, we assume the transition probability between all states are equal-probable.

5.2. Selected techniques for comparison

We compare the proposed technique with a time domain MAP estimator [4] and a binary mask based separation technique [2, 5]. The time domain MAP based estimator is expressed by

$$\{\hat{s}_1^t, \hat{s}_2^t\} = \arg\max_{\mathbf{s}_1^t, \mathbf{s}_2^t} p_{\mathbf{z}^t|\mathbf{s}_1^t, \mathbf{s}_2^t}(\mathbf{z}^t|\mathbf{s}_1^t, \mathbf{s}_2^t) p_{\mathbf{s}_1^t}(\mathbf{s}_1^t) p_{\mathbf{s}_2^t}(\mathbf{s}_2^t).$$

It is proven that [4] this maximization leads to Wiener filtering in the frequency domain. More specifically, for speaker one we have

$$\widehat{\mathcal{F}_D\big(s_1^t(d)\big)} = \frac{S_1^{i^*}(d)}{S_1^{i^*}(d) + S_2^{j^*}(d)} \mathcal{F}_D\big(z^t(d)\big), \quad d = 1, \ldots, D$$

where $S_1^{i^*}(d)$ and $S_2^{j^*}(d)$, $d = 1, \ldots, D$ denote the power spectral densities corresponding to the state $y_1^{i^*}$ of speaker one and state $y_2^{j^*}$ of speaker two, respectively. Note that, as opposed to the model introduced in [4], we assume that the model is free of noise, that is $z^t(d) = s_1^t(d) + s_2^t(d)$. The states $y_1^{i^*}$ and $y_2^{j^*}$ whose power spectral densities are used in the Wiener filter are obtained from

$$\{y_2^{j^*}, y_1^{i^*}\} = \arg\max_{y_1^i, y_2^j} \ln \prod_{d=1}^{D} p_{z|y_1^i, y_2^j}\big(|\mathcal{F}_D\big(z^t(d)\big)|^2 | y_1^i, y_2^j\big) \times p_{y_1^i}(y_1^i) p_{y_2^j}(y_2^j) \quad (18)$$

where $p_{z|y_1^i, y_2^j}\big(|\mathcal{F}_D\big(z^t(d)\big)|^2 | y_1^i, y_2^j\big)$, $p_{y_1^i}(y_1^i)$, and $p_{y_2^j}(y_2^j)$ are, respectively, the frequency domain counterpart of $p_{z|x_1^i, x_2^j}\big(z(d)|x_1^i, x_2^j\big)$, $p_{x_1^i}(x_1^i)$, and $p_{x_2^j}(x_2^j)$ defined in the log frequency domain. It is should be noted $\widehat{\mathcal{F}_D\big(s_2^t(d)\big)}$ is obtained in a similar fashion to $\widehat{\mathcal{F}_D\big(s_1^t(d)\big)}$ with only swapping i and j as well as the subscripts 1 and 2.

The second technique chosen for comparison is the binary mask method which is a log frequency domain technique like the MAP based separation technique proposed in this paper. Using the same notations applied for the MAP based estimator, the estimate of the sources using the binary mask can be expressed by

$$\hat{s}_1(d) = \Omega\big(\mu_1^{i^*}(d) - \mu_2^{j^*}(d)\big) z(d), \quad \text{and}$$
$$\hat{s}_2(d) = \Omega\big(\mu_2^{j^*}(d) - \mu_1^{i^*}(d)\big) z(d) \quad d = 1, \ldots, D$$

where i^* and j^* are obtain using (16). In the next subsection, we report the objective results obtained from the MAP-based technique, Wiener filter, and the binary mask.

5.3. Experimental results

The estimated frames of each speech file are multiplied by the phase of the observation signal and transformed to the time domain using the inverse D-point Fourier transform. The resulting frames are multiplied by a Hanning window and then the overlapped-add method is used to reconstruct the estimated speech files. In order to evaluate the quality of separated speech files, SNR is used as a objective measure.

The SNR results for the separated male-male, female-female female-male mixtures are listed in Tables 1, 2, and 3, respectively. From the tables, we observe that, on average, the MAP based separation technique in the log frequency domain outperforms the two other techniques. It is also seen the worst SNR results are obtained from the time domain MAP estimator.

Table 1: SNR (dB) results obtained from MAP log domain, Binary mask, and MAP time domain for five male-male mixtures. The last row presents the averaged SNR for each technique.

Mixture	Gender	Log Frequency MAP	Binary Mask Filter	Time Domain MAP
1	m	5.53	**5.73**	5.48
	m	**5.93**	5.29	5.00
2	m	4.26	**4.31**	4.20
	m	**5.37**	4.73	4.76
3	m	**5.33**	5.10	3.24
	m	**4.54**	4.01	2.45
4	m	4.33	3.29	**4.35**
	m	3.57	1.85	3.05
5	m	5.59	**5.61**	5.30
	m	**5.71**	4.77	4.69
	Ave	**5.01**	4.46	4.25

Table 2: SNR (dB) results obtained from MAP log domain, Binary mask, and MAP time domain for five female-female mixtures. The last row presents the averaged SNR for each technique.

Mixture	Gender	Log Frequency MAP	Binary Mask Filter	Time Domain MAP
1	f	**8.77**	7.53	8.26
	f	**8.45**	7.19	7.95
2	f	**7.48**	7.01	6.72
	f	**7.68**	7.03	6.83
3	f	9.55	**9.57**	9.16
	f	**9.72**	9.55	9.13
4	f	6.24	4.33	**6.38**
	f	6.40	4.54	**6.61**
5	f	**6.20**	5.63	6.16
	f	**6.17**	5.58	6.10
	Ave	**7.66**	6.79	7.33

6. Conclusions

In this paper, using maximum a posteriori estimation approach a two state, non linear filter was derived to estimate the log spectra of the sources from the mixture. It was shown that the derived filter outperforms the binary mask and Wiener filters which have been extensively used in source separation techniques. The proposed filter can be favorably used in model-based single channel techniques where *a priori* knowledge of sources are available.

Table 3: SNR (dB) results obtained from MAP log domain, Binary mask, and MAP time domain for five female-male mixtures. The last row presents the averaged SNR for each technique.

Mixture	Gender	Log Frequency MAP	Binary Mask Filter	Time Domain MAP
1	f	**7.31**	6.64	6.31
	m	5.88	**6.00**	5.67
2	f	8.53	**8.73**	7.94
	m	**8.19**	7.88	7.20
3	f	**10.50**	9.35	8.42
	m	8.16	**8.87**	8.01
4	f	**9.76**	8.42	6.40
	m	**8.79**	8.04	6.08
5	f	**8.67**	8.64	7.42
	m	**8.60**	8.30	7.13
	Ave	**8.42**	8.00	7.10

7. References

[1] D. Ellis, "Model-based scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G. Brown, Eds. Wiley/IEEE Press, 2006.

[2] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "Performance evaluation of three features for model-based single channel speech separation problem," in *In Proc.ICSLP06*, Sept. 2006.

[3] S. T. Rowies, "Factorial models and refiltering for speech separation and denoising," in *EUROSPEECH–03*, vol. 7, May 2003, pp. 1009–1012.

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

[5] S. Roweis, "One microphone source separation," in *Proc. Neural Inf. Process. Syst.*, 2000, pp. 793–799.

[6] T. Virtanen, "Speech recognition using factorial hidden markov models for separation in the feature space," in *In Proc. ICSLP06*, Sept. 2006, pp. 173–176.

[7] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *In Proc. ICSLP06*, Sept. 2006.

[8] G. J. Jang and T. W. Lee, "A probabilistic approach to single channel source separation," in *Proc. Advances in Neural Inform. Process. Systems*, 2003, pp. 1173–1180.

[9] A. M. Reddy and B. Raj, "A minimum mean squared error estimator for single channel speaker separation," in *INTERSPEECH–2004*, Oct. 2004, pp. 2445–2448.

[10] 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," *Electronic Letters*, vol. 42, no. 12, pp. 75–76, June 2006.

[11] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *JASA*, Nov. 2005, http://www.dcs.shef.ac.uk/spandh/gridcorpus.