Implementation of Frequency Domain Approach Using Instantaneous Mixing Auto Recursive for Separation of Speech Signals

C. Anna Palagan and K. Parimala Geetha

ECE, Rajas Engineering College, Vadakkangulam, Tirunelveli, Tamilnadu, India
ECE, Ponjesly College of Engineering, Nagercoil, Tamilnadu, India

Abstract: In the present work a novel algorithmic rule by taking the speech from two different microphones and separate these speeches by prediction of separating speech mixtures that is predicated on separation matrices is planned. In multi-talker applications so as to boost individual speech sources from their mixtures is done by Blind source Separation (BSS) ways. From the previous published works of separation of speech signals, the main disadvantage is that the incidence of distortion present within the signal that affects separated signal with loud musical noise. The idea for speech separation in standard BSS ways is simply one sound source in a single room. The proposed methodology uses as a network that has the parameters of the IMAR model for the separation matrices over the complete frequency vary. An attempt has been made to estimate the best values of the IMAR model parameters, \( \Phi_w \) and \( \Phi_v \), by suggests that of the maximum-likelihood estimation methodology. Based on the values of these parameters, the source spectral part vectors are estimated. The entire set of TIMIT corpus is employed for speech materials in evolution results. The Signal to Interference magnitude Relation (SIR) improves by a median of 6dB sound unit over a frequency domain BSS approach.

Key words: Blind Source Separation (BSS) - Separation Matrices - Instantaneous Mixing plus Auto Regressive (IMAR) Model - Maximum Likelihood Estimation

INTRODUCTION

The audio and speech signal processing applications, the separation of speech signals is very important done by using Blind Source Separation (BSS) technique. The BSS has been used in multi talker applications and acoustics signal processing. Several adaptive step size methods for BSS for Robot audition systems have been proposed. The parameters are not adjusted manually and there is no need of additional preprocessing. For the moving sources [1-19] the positions and velocities of the source is obtained from the 3-D tracker based on Markov Chain Monte Carlo particle filter which results in high sampling frequency. In this method it provides separation of the sources without any prior knowledge of moving sources and it is used to perform real time speech enhancement. In frequency domain, the permutation problems [13] a new technique which partitioned the full frequency bands into small regions by using correlation of separate signal powers. The region wise permutation alignment is performed by region growing manner. Here the permutation alignment is based on inter frequency dependence of separated power signal.

For multi channel acoustic echo cancelation [8] the ICA is jointly perform source separation and multichannel acoustic cancellation through semi BSS without double track detection. To reduce the effect of non uniqueness the matrix constraint is used. For detecting a time varying mixing matrix the short time Fourier transform [20] is used. In frequency domain [13, 6] for reducing the computational complexity and increase the speed frequency domain BSS is used. In this algorithm the higher order frequency dependencies which employing real conference room recordings. For extracting independent from array signals the Bi iterative algorithm [3] is used. For solving the non unitary joint diagonalization problem in BSS a simultaneous Bi-iterative algorithm is introduced. In the multiple sources case, in order to find the accurate estimation of propagation time delays [7]. Generalized State Coherence
Transform (GSCT) which is a non-linear transform of the space represented by the whole demixing matrices. In convex divergence ICA algorithm, [11] the source signals of the blind sources is derived by the characteristics of Parzen window based distribution. Based on the experimental results this algorithm is the fast algorithm for the blind sources which involves speech and music signals. For cancelling the echo’s in the blind sources during continuous double track [10] in order to estimate the blind source signals the maximum likelihood approach is used.

Blind source separation (BSS) methods aim to achieve this goal, based on some prior knowledge of the source signal properties. Following the physics of sound mixing, let us consider N sources \( s_m(t) \), \( m = 1, \cdots, N \), to be convolutively mixed. At M sensors, the recorded mixture signals \( a_i(t) \), \( i = 1, \cdots, M \), is denoted by

\[
a_i(t) = \sum_{m=1}^{N} \sum_{d=1}^{L} g_{m}(d) s_{m}(t - d)
\]

where \( L \) is that the delay length on the order of \( 10^3 \) faucets (each faucet last \( 1/F_s \) second wherever \( F_s \) is that the sampling frequency) in an exceedingly commonplace area \( g_{m}(d) \) is that the separate Green’s perform of the area, jointly called the area impulse response (RIR). The (severely ill-posed) mathematical drawback is to recover each \( G_{m}(d) \) and \( s_{m}(t) \) from \( A_{i}(t) \). A serious branch of BSS is that the thus referred to as freelance element analysis (ICA) that assumes that the supply signals square measure orthogonal to (or freelance of) one another [2].

ICA is a more general methodology than recovering sound signals. The time domain ICA [1, 5] attempts to estimate the \( g_{m} \)’s directly in order to deal with a high dimensional nonconvex optimization problem [2]. Frequency domain ICA [15] & [16] solves an instantaneous (\( L = 0 \)) version of (1) in each frequency bin after applying the discrete Fourier transform (DFT) to (2) frame by frame:

\[
A_i(f, \tau) = \sum_{m=1}^{N} G_{m}(f) S_{m}(f, \tau)
\]

where \( (A_{i}, G_{m} \& S_{m}) \) square measure the \( T \)-point DFT of \( (a_{i}, g_{m} \& s_{m}) \) severally and \( \tau \) is that the frame variety. The larger \( T/l \) square measure, the higher the approximation. Because of the absence of the regularity in \( d \) of \( g_{m} \) and \( s_{m} \), DFT doesn't rework convolution to native product precisely. The frequency domain approach is limited to use a long DFT. In addition to computations to sort out scaling and permutation ambiguities when synthesizing multi-frequency estimation of \( S_{m}(f, \tau) \) back to a time domain output [2, 14]. Imperfections and errors in scaling and permutation in the frequency domain may lead to artifacts in the time domain signals at the final output.

**Existing Methods and New Idea:** BSS refers to the problem of recovering signals from several observed linear mixtures. Up to now, solving the BSS problem in an underdetermined case has mainly consisted in assuming that the speech signals were sufficiently sparse [22, 21], However, due to unexpected discontinuous zero-padding, such separated signals have considerable distortion and therefore a loud musical noise is heard. In [4], an estimated mixing matrix was used for solving the determined BSS problem. Our suggestion for eliminating the distortion matter is to combine sparseness with mixing matrix estimation. Indeed we can obtain more information about the signals to be separated and to reduce the zero-padding effect, from which the musical noise originates. Whereas Vielva et al., Rickard and Yilmaz worked on an underdetermined instantaneous case employing sparseness [22, 21, 4] and Deville on a determined instantaneous case utilizing a mixing matrix estimation [4], here, we are dealing with underdetermined BSS in a convolutively case.

**Blind Source Separation of Speech and Music Signals:**

Blind source separation (BSS) could be a technology for separating mixtures of multiple speech signals. This technology has been studied extensively and vital progresses are modeled the last decade. However, typical BSS ways performs terribly poor once the reverberation time is massive. Many researchers have addressed this downside, however it's still associate open question. Our approach to overcoming this limitation is to unify BSS and performing BSS will realize BSS even under highly reverberant environments.

Figure 1 is an example for speech separation of an indoor auditorium problem. For these applications, the instant combination model could also be applicable as a result of the propagation delays is negligible. However, in real environments substantial time-delays could occur in a design and algorithmic rule is required to account for the blending of time-delayed sources and convolved sources. It focuses on the implementation of the learning algorithm and on issues that arise when separating speakers in room recordings. It used associate approach during a feed forward network enforced
Proposed Method

Imar Model to Generate Microphone Signals: In the proposed method the BSS is recovered of source signals by LTI filter and permutation. The time domain Blind Source Separation approach is used here. In this proposed method the estimation of blind source signal is in the form of source signal vector $B_i(n)$ by applying an $I_{in}$ input signals and $I_{out}$ output separation filter value to observed signal vector $O(n)$. In the time domain BSS approach for separating sound mixtures the order of the separation filter is set a value that exceeds the room reverberation time. The order of the separation filter becomes very large for the reverberation time is long. So the convergence rate is poor and the cost for computation is very high. The estimation of source spectral component vector in the frequency domain BSS approach is done by applying a separation matrix to the observed spectral component vector.

In this proposed method the assessment is a multiple sound source case, where $I_{in} = 2$. In this consideration the frequency domain BSS approach as shown in Figure 2 is by using WPE method as a preprocessor which illustrate the case of $I_{in} = I_{out} = 2$. Because in the first step we use prediction error with first microphone for BSS process. Figure 3 shows the structure for Mixing of Speech signals.

For any microphone $m$ the prediction target as

$$P_{n,m} = O_{n,m} - \sum_{s=L_1}^{L_2+M_1-1} h_{n,s,m} O_{n,s} x_{n} \leq n \leq I_{M}$$

(3)
where \( h_{n,s} \) \( L = L + M -1 \) denoted the prediction filter for the \( i \)th microphone spectral component and \( P_{n,s} \) is the corresponding prediction error. The different spectral component outputs \( P_{i,s} \ldots P_{n,s} \) can be obtained. The instantaneous mixtures of the source spectra components were considered for these components. Based on the theory of Multichannel linear prediction the values of \( P_{i,s} \ldots P_{n,s} \) become nearly instantaneous mixtures by using appropriate prediction filters although such prediction filters may not be able to obtained with the WPE method. For the \( m \)th microphone the prediction filter values are \( D_{n,s} \).

It is assumed that the bin indices is 1 for all frequencies from the set of exists values of \( X \) and \( \{ D_{n,s} \} \) \( L - L + M -1 \)that is equalize the output spectral component vector \( B_{s,v} \) based on these assumptions we identified \( Z_{i,n} \) with \( B_{s,v} \) is

\[
Q_{n,v} = \sum_{s=L}^{L + M -1} H_{s,v} O_{n,s,v} + P_{n,v}
\]

where \( \{ h_{n,s} \} \) \( L = L + M -1 \) denoted the prediction filter for the \( i \)th microphone spectral component and \( P_{n,s} \) is the corresponding prediction error. The different spectral component outputs \( P_{i,s} \ldots P_{n,s} \) can be obtained. The instantaneous mixtures of the source spectra components were considered for these components. Based on the theory of Multichannel linear prediction the values of \( P_{i,s} \ldots P_{n,s} \) become nearly instantaneous mixtures by using appropriate prediction filters although such prediction filters may not be able to obtained with the WPE method. For the \( m \)th microphone the prediction filter values are \( D_{n,s} \).

It is assumed that the bin indices is 1 for all frequencies from the set of exists values of \( X \) and \( \{ D_{n,s} \} \) \( L - L + M -1 \)that is equalize the output spectral component vector \( B_{s,v} \) based on these assumptions we identified \( Z_{i,n} \) with \( B_{s,v} \) is

\[
Q_{n,v} = \sum_{s=L}^{L + M -1} H_{s,v} O_{n,s,v} + P_{n,v}
\]

The assumption taken in the equation (4) will not completely hold in real time so further experimental detail is desired. So in IMAR model it performs high separation of speech signal based on the possible assumption is at least partially show the practical validity of this assumption. Figure 4 shows the filter representation of mixing demixing of speech signals.

The set of equations (3) and (4) represents the generation of IMAR model for the observed spectral component vector \( O_{n,v} \). In need of this mode may be interpreted as follows. The individual sound source signals and the spectral component of the sound source was given in equation (4) are instantaneously mixed together with mixing \( X_{i,v}^{-1} \) to form \( P_{u,v} \). In the equation (3) the mixing of the remaining elements present in \( P_{u,v} \) with the multichannel AR system with regression or prediction matrices \( \{ H_{i,n} \} = L - s + L - s -1 \) to generate the observed spectral component vector \( O_{u,v} \).

Figure 5 represents the Instantaneous Mixing plus Auto Regressive (IMAR) model. The latent spectral component vector is unobservable by \( P_{u,v} \). The separation
matrices and prediction matrices parameters of IMAR model over the entire frequency range is denoted by $\Phi_w$ and $\Phi_v$. Figure 6 shows the analysis filter for instantaneous mixing plus auto regressive model. The following expressions represent the set of separation matrices and prediction matrices as

$$\Phi_w \{ W_v \}; \quad 0 = v = K-1$$

$$\Phi_v = \{ D_v \} L_i \leq s \leq L_v + S_v^{-1}; \quad 0 \leq v \leq K-1$$

By using the maximum likelihood estimation the ideal values of IMAR model parameters $\Phi_w$ and $\Phi_v$. The parameter values of the source spectral component vectors are estimated from equations (3) and (4). The matrix form is represented for the IMAR model in the MIMO filter which gives the relationship between the IMAR model and the frequency domain. The expression for matrix form of $X_{sv}$ as

$$X_{sv} = \begin{bmatrix} X_s & \text{if} & s = 0 \\ 0 & \text{if} & 1 \leq s \leq L_s \\ D_{sv}X_v & \text{else} \end{bmatrix}$$

where $0$ is $z_{sv} - \sum_{i=s+1}^{L_v} X_vO_{sv}$. A zero matrix. Then (3) and (4) may be summarized in one equation as

Parameters Estimation Based on Maximum-likelihood Function: The IMAR mode parameters $\Phi_w$ and $\Phi_v$ was optimized by using the maximum log likelihood function. For finding the source spectrum IMAR model first find out the time varying all pole model. By using the time varying all pole and IMAR model we derive the log likelihood function. The set of Linear Predictive Coding and PRPs over the whole observation period we use $\Phi$, and it is represented as

$$\Phi = \{ \mu_{m,v}, \sigma_{m,v}^2 \}; \quad 1 \leq m \leq M, \quad 0 = t = N - 1$$

The Linear Predictive Coding and PRPs are collectively called the all pole parameters. It is assumed that, the speech signal is taken as short time based signal which has a short time frame is 18ms. So that in the present works the short time analysis of frame length 18ms. In added with this the smallest frame length less than 18ms is not considered because its violates the assumption.

The source spectrum parameters of both LPCs and PRPs can be observed by means of probability distribution function is $H = \{ h_v \}; \quad 0 = u = M-1, \quad 0 = l = D-1$. The probability density function for the source spectral component vector $k_{uv}$ by the assumptions of $h$ (2) and $h$ (3) can be written as

$$p(k_{uv}; \Phi) = \mathcal{N}( k_{uv}; 0; kA_{uv})$$

where $0$ is a zero vector $kA_{uv}$ is the diagonal matrix defined as

$$kA_{uv} = \begin{bmatrix} kA_{uv} & 0 \\ 0 & kA_{uv} \end{bmatrix}$$


Fig. 6: Analysis filter for Instantaneous Mixing plus Auto Regressive model
The $P_u$ is the probability distribution function of the latent spectral component vector which is taken by the equations (5) and (11).

$$p(P_u; \Phi_s; \Phi_w) = N_c \{P_u; 0, pA_x \}$$  \hspace{1cm} (12)

where the covariance matrix $pA_x$ is given by

$$pA_x = (X_{A}X_{A})^{-1}$$  \hspace{1cm} (13)

The past sequence of the probability function for observed spectral component vector is taken from the equations (5) and (13) and it is represented by

$$p(O_{w+1} \ldots O_{w+L-1}; \Phi_s; \Phi_w; \Phi_G) =$$

$$N_c \left\{ O_{w+1}; \sum_{s=2}^{L_N} H_{w,s}^{(w)} O_{w+1,s}; pA_{x,s} \right\}$$  \hspace{1cm} (14)

The probability density function for observed data is given by

$$P(O; \Phi_s; \Phi_w; \Phi_G) = \sum_{v=0}^{D_N} \sum_{u=0}^{D_N} p(O_{v+1} \ldots O_{v+L-1}; \Phi_s; \Phi_w; \Phi_G)$$  \hspace{1cm} (15)

The maximum log likelihood function was obtained by taking the log value of equation (15).

$$l(\Phi_s; \Phi_w; \Phi_G) = \sum_{v=0}^{D_N} \sum_{u=0}^{D_N} \left[ \log pA_{x,s} \right] -$$

$$\left( O_{w+1}; \sum_{s=1}^{L_N} H_{w,s}^{(w)} O_{w+1,s} \right)^{(w)}$$  \hspace{1cm} (16)

From the equations [1-16] it is identified that the maximum log likelihood function not depends on the separation matrices $\Phi_w$ and prediction matrices $\Phi_G$ but also depend on the pole parameter $\Phi_s$. So the maximum log likelihood function is calculated which depends on $\Phi_s$, $\Phi_w$ and $\Phi_G$ is used as the estimates of the ideal values of $\Phi_w$ and $\Phi_G$, respectively.

**RESULTS AND DISCUSSIONS**

In the proposed method two sources and two microphones are used for testing the speech signals. The complete test used a set of TIMIT corpus which includes 38 male speakers, 8 female speakers and 145 utterances. The sampled frequency used for the testing of acoustic signals of these utterances is 14 KHz and the bandwidth is limited between 70 Hz to 4 KHz frequency range. The data’s from 26 male speakers are taken and it can be used from male – male utterances pairs and the data’s from 5 female speakers are taken and it can be used for female – female utterances pairs. The remaining data’s from 12 male speakers and 3 female speakers are taken and it can be paired to form the remaining utterances. So in total 55 male – male, 20 female – female and 70 male – female utterances pairs were generated. We take each utterances pair and it can be mixed with the acoustic signals of the two utterances with the room impulse responses measured in a varichoic chamber to simulate signals which can be observed from the microphones.

Figure 7 shows an experimental setup developed for the present work. The Signal to Interference Ratio (SIR) and the Direct to Reverberation Ratio (DRR) should be evaluated from each trial taken by the experimental setup. The $I_m$ source from the $I_m$th microphones are described by the following equations.

![Fig. 7: Experimental setup](image-url)
The value of early reflection delay $\Delta$ was set as 40 ms. The signal to interference ratio and the direct to reverberation ratio output also defined in the same way. From the experimental output each output signal is decomposed according to the sources in order to find the values of output signal to interference ratio. Let us consider $W_s(n)$ denote the output signal $I_s^n$ and $W_{is}^n(n)$ denote the $I_s^n$ source component. For these calculations of processing microphone signals $O_{is}^n(n)$, $O_{is}^n(n)$ with the estimated prediction and separation matrices. For calculating the direct to reverberation ratio we need the impulse responses from the sources to the outputs. For estimating the impulse responses it uses the least mean squares matching. The impulse response is estimated from the source speech signal $I_s$ to the output speech signal $I_o$. The matching errors are taken by the assumptions of -20 dB. This experimental result indicates the IMAR model is effective for BSS.

Figure 8 shows the input samples of speech signals from two sources. The corresponding short time Fourier transform was illustrated in Figure. 9. The instantaneous mixing of the samples based on the IMAR model was shown in Figure. 10. The separated speech signals of the blind sources are illustrated in different forms for various separable speech signals are given in Figure. 11.

Figure 12 and Figure. 13 illustrated the average variations SIRs and DRRs based on different reverberation. From that figure we can clearly mentioned the IMAR model gives better outperformed in the frequency domain blind source separation for both reverberation conditions in terms of average SIR. Figure 14 shows the separated speech with echo canceller by using IMAR for $\mu=0.025$ and 0.04. Figure 15 represented the separated speech with noisy environment.

Figure 16 shows the average SIRs for each reverberation condition. We infer that the methodological difference between the ICA, CSD and IMAR models leads to the difference in SIR improvement.

![Fig. 8(a): Input Signal](image1)

![Fig. 8(b): Noise Signals](image2)
Fig. 9(a): Input Signal + Noise Signal

Fig. 9(b): Speech Signals from Mics

Fig. 10(a): Speech Signals from mic

Fig. 10(b): Speech Signals from mic

Fig. 11(a): Speech Signals from mics

Fig. 11(b): Noisy Reverberant Signal

Fig. 12(a): Reverberation and Echo Signal

Fig. 12(b): Speech value from Mic
Fig. 13: Estimated Speech Signal

Fig. 14: Separated Speech with Echo Canceller by using IMAR for $\mu=0.025$ and 0.04

Fig. 15: Separated Speech with noisy environment
Table 1: Average SIR and DRR Increases In Decibel

<table>
<thead>
<tr>
<th>Gender for Speaker 1</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIR</td>
<td>Speaker 1: 5.21</td>
<td>4.72</td>
</tr>
<tr>
<td></td>
<td>Speaker 2: 5.26</td>
<td>4.88</td>
</tr>
<tr>
<td></td>
<td>Speaker 3: 5.30</td>
<td>4.92</td>
</tr>
<tr>
<td>DRR</td>
<td>Speaker 1: 4.32</td>
<td>4.12</td>
</tr>
<tr>
<td></td>
<td>Speaker 2: 4.51</td>
<td>4.42</td>
</tr>
<tr>
<td></td>
<td>Speaker 3: 4.68</td>
<td>4.58</td>
</tr>
</tbody>
</table>

Table 2: Comparison of SIR Improvement Between Different Algorithms and the Proposed Method

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>SIR Improvements in dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICA</td>
<td>03</td>
</tr>
<tr>
<td>CSD</td>
<td>04</td>
</tr>
<tr>
<td>IMAR</td>
<td>6.5</td>
</tr>
</tbody>
</table>

From the above results and comparative tables we proved that IMAR model is the best one which is used for the applications of separation of speech signals. Figure 17 shows the comparison of DRR between ICA, CSD by using IMAR model [17-22].

CONCLUSION

The present work is carried out to design the effectively separate of speech signal from the blind Source Separation by using the method of Instantaneous Mixing Auto Regressive method and the maximum likelihood function. The key features presented in the Instantaneous Mixing Auto Regressive method is that optimized separation of speech signals and thereby enabling us to perform a blind source separation process in consideration. In the present method the signal to interference rate improves over 6 dB. By using Instantaneous Mixing Auto Regressive method it attained good signal to interference ratio and direct to reverberation ratio even when a reverberation time was 0.3 s. It is concluded that the Instantaneous Mixing Auto Regressive method provides a powerful tool for microphone array signal processing in a reverberant room impulse response.
REFERENCES