Undetermined Convolutive Blind Source Separation

Sugumar D
Karunya University
Coimbatore

Sindhu Ann John
Karunya University
Coimbatore

P.T Vanathi, PhD.
P.S.G Tech
Coimbatore

ABSTRACT

This paper presents a blind source separation process for convolutive mixtures of audio sources. Here undetermined condition that is few microphones than sources has been considered as a mixing model. By an expectation–maximization (EM) algorithm the separation operation is performed in the frequency domain. The T-F masking separation is made use which is a powerful approach for the separation of underdetermined mixtures, especially for the separation of single-channel mixtures. Even under reverberant conditions the process enables to attain a good separation. From the experimental results, separated signals SDR values of speech mixtures is obtained in the range of 7.5dB while for music mixtures in the range of 2.9dB. It can be concluded from these values that separation of speech mixtures is better than music mixtures.

General Terms

Blind Source Separation, Convolutive mixing

Keywords

Blind Source separation, Convolutive mixtures, undetermined mixtures, EM algorithm, T-F masking.

1. INTRODUCTION

Blind Source Separation (BSS) is the separation of a set of signals from a set of mixed signals without the aid of information (or with very little information) about the source signals or the mixing process [1]. The term blind refers to only that the mixture is known to us.The assumption that the source signals do not correlate with each other is relied upon for the separation. Blind Source Separation refers to two classes of multichannel signal processing tasks in which the goal is to extract multiple useful signals from the convolutive The mixing model is shown in Fig.1.The mixtures are assumed to be linear superposition of source signals. The two types of mixing are instantaneous and convolutive mixing. The instantaneous mixture is given as

$$x(n) = As(n) + v(n) \qquad (1)$$

where A is an M× N mixing matrix, n denotes the discrete time index and v(n) is additional noise.The convolutive mixture is given as

$$x(n) = \sum_{k=0}^{k-1} A_1\, s(n-k) + v(n) \quad (2)$$

where A_1 is an N×K mixing matrix. The convolutive mixture is more applicable for separation of speech signals because the convolutive model takes reverberations into account. The separation of convolutive mixtures can either be performed in the time or in the frequency domain.There are different methods of blind source separation like Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Independent Component Analysis (ICA) and Dependent Component Analysis (DCA).

The majority of existing techniques rely on time-difference-of-arrival for each source at multiple microphones [6],[7]. The separation process employs a widely used T–F masking scheme to separate the mixtures into individual sources. In the two-stage approach adopted, the first stage is responsible for frequency bin-wise clustering. To group together the bin-wise separated frequency components coming from the same source an additional task is performed in the second stage which is almost identical to the permutation problem of frequency-domain ICA-based BSS [2],[3].In the proposed method, the bin-wise clustering results of the first stage are represented by a set of posterior probabilities. BSS for convolutive mixture can be solved efficiently in the frequency domain, where ICA is performed separately in each frequency bin[3In the second stage, the mixtures are separated using the estimated mixing matrix from the first stage [4].

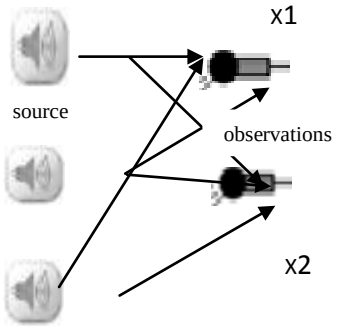

Fig1.Undetermined convolutive Mixing Model

A related difficult problem is determining in which regions of a spectrogram a sound is observable. It is exacerbated by the presence of sounds from other sources and by realistic reverberations, as would be found in a classroom[5]. A two-step method based on a general maximum a posteriori (MAP) approach was adopted. In the first step, the mixing matrix was estimated based on hierarchical clustering, assuming that the source signals are sufficiently sparse. The algorithm works directly on the complex-valued data in the time-frequency domain and shows better convergence than algorithms based on self-organizing maps.In this paper we are considering the separation of undetermined sources using bin-wise clustering in T-F domain.

This paper is organized as follows. Section II provides a system overview of the proposed method. Sections III and IV presents detailed explanations of the first and second stages of the proposed method, respectively and the experimental results. Section V concludes this paper.

2. SYSYTEM OVERVIEW

This section provides a system overview of the proposed BSS Method. Fig. 2 shows a processing flow for T–F masking based BSS. Pre-processing is important while taking audio signal as the input signal. Pre-processing contains the following procedures:

- Mixing
- Framing
- Masking
- STFT

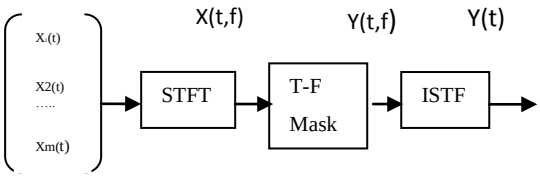

Fig.2.Processing flow of BSS with time–frequency masking

2.1 Sources

As shown in Fig. 3, let s1, s2 and s3 be the source signals and X1,...,XM be microphone observations. The numbers of sources and microphones are denoted by N and M, respectively. The number of sources can be greater or lower than the number of microphones. A case where N>M is called an underdetermined BSS, and alternatively a case where N<M is called overdetermined BSS and N=M is determined BSS.

The observation xj at microphone j is described by a mixture

$$Xj(t) = \sum_{k=1}^{N} Sjk(t) \qquad (3)$$

of source Sk images at the microphone j

$$Sjk(t) = \sum_{l} hjk(l)Sk(t-l) \qquad (4)$$

wheret represents time and hjk represents the impulse response from source k to j microphone . The goal of BSS is

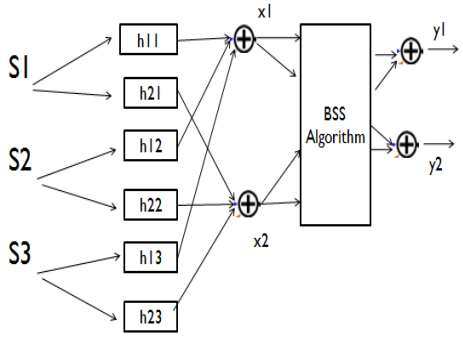

Fig.3.Signal notations

to obtain sets of separated signals {y11....y1M},...,{yN1.....yNM} where each set corresponds to each of the source signals .Various source combinations are available with male and female sources as shown in Table1. The number of sources taken should satisfy the undetermined case.The experimental specification for sources is indicated in TableI.In this paper, two speechesS1, S2 and a music sourceS3 are considered with same source specification as tabulated.

Table 1.Source Specification

Sources	Type	No. Of Samples	Duration(sec)	Sampling Frequency (Hz)
S1	Speech	118976	6	16000
S2	Speech	118976	6	16000
S3	Music	118976	6	16000

2.2 Mixing

Mixtures of audio sources can be acquired in many ways, having the recording set-up a big influence on the different separation approaches. In this context, synthetic audio mixtures, which are artificially created, are very different in nature from real recordings.The recordings are obtained by capturing several sources that have been physically mixed by the arrangement as shown in Fig.7.The computerized recording set up for undetermined convolutive mixtures with linear source-microphone (S-M) arrangement is shown in Fig.4 and 5. In the case of real recordings, the room where the sound is recorded plays also an important role, as the recorded signals are the total contribution of direct path signals from the sources and the reflections that occur inside the room. Reverberation is usually characterized by the room reverberation time, RT60.

Fig.4. Linear mic set up

Fig.5. Computerized set up for recording

In the experimental set up, three sources are considered. The arrangement of S-M is in the form of linear or circular as shown in Fig6and Fig7.The S-M arrangement may or may not have LOS with minimum /maximum delay with sources as in Fig7. The same set up is used in this paper.

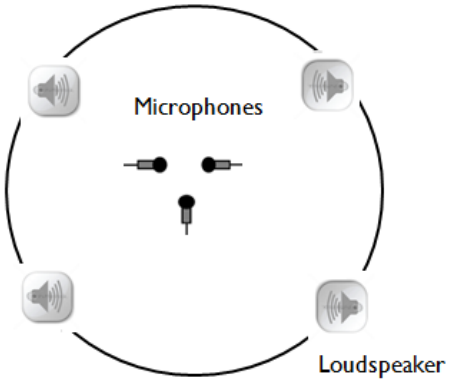

Fig.6. Circular arrangement of mixing with 4 speakers and three microphones

For this paper, the number of sources taken is three and the number of microphones two. The mixing arrangement gives two mixtures of specification as in Table2.

Table 2. Mixed signal specification

Mix	Length	Time (sec)	Sampling Freq(KHz)
Mix 1	122576	6	16
Mix 2	122576	6	16

2.3 Framing

Framing refers to the process of partitioning an audio into multiple segments. The goal of segmentation is to simplify and/or change the representation of an signal into something that is more meaningful and easier to analyze.

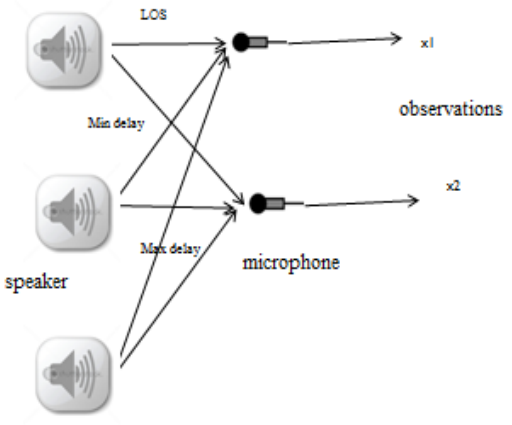

Fig.7 Linear arrangement of mixing with three speakers and two microphones

Let x(t) be a continuous signal which is to be sampled, and that sampling is performed by measuring the value of the continuous signal every T seconds, which is called the sampling interval. Thus, the sampled signal x[n] given by:

$$x[n] = x(nT) \qquad (5)$$

With n = 0, 1, 2, 3, ... The microphone observations sampled at a sampling frequency fs are converted into frequency-domain time-series xj(τ,f) signals by an STFT with an -sample frame.

$$\sum_0^l xj(\tau, f) = win(t')xj(t' + \tau)e^{-2\lambda ft'} \quad (6)$$

For frame indices τ =0,St_s,...(T-1) and frequencies f=0,(1/L)fs,...,((L-1)/L)fs. The analysis window $win_a(t)$ such as Hanning and Hamming window tapers smoothly to zero at each end.

$$wina(t) = (1/2)(1 - cos(2\Pi t/Lts) \quad (7)$$

The frame size used is 1024(8khz) or 2024(16khz).

2.4 Short Time Fourier Transform

One approach whichgive information on the time resolution of the spectrum is the short time fourier transform (STFT). In STFT, the signal is divided into small enough segments, which can be assumed to be stationary. For this purpose, a window function w is chosen. The window function is first located to the very begging of the signal otherwise the window function is located at t=0.

The microphone observation sampled at a sampling frequency fs, or with a sampling period ts=1/fs, are converted into frequency-domain time-series signals Xj(ι,f) by an STFT with an -sample frame and sample shift.

$$\text{Xj}(\iota, \text{f}) = \sum_{t=0}^{.(L-1)ts} wina(t\ \ ')Xj(t' + \iota)e^{-2\pi ft} \quad (8)$$

for frame indices *ι=0*,St_s,...,T-1 and frequencies f=0,(L-1)fs,....,((L-1)/L)fs.

2.5 Inverse Short Time Fourier Transform

To reconstruct the original files back Inverse Short time FourierTransform (ISTFT) is performed.The STFT is invertible and the original signal is recovered fromISTFT. Time-domain separated signals are calculated with an inverse STFT applied to the separated frequency components where the summation over frequencies and the summation over frame time indices satisfy the conditions.

3. SOURCE SEPARATIONVIA BIN-WISE CLUSTERING

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense) to each other than to those in other clusters.The clustering process fine tunes the separation with proper alignment.

To determine the Gaussian or non- Gaussian nature of signals parameters such as: Mean, Variance, Skewness, Kurtosis, Negantrophy are calculated. In case of skewness, the value can be positive or negative, or even undefined.

Table 3.Statistical parameters

Mix	Mean	Variance	Skewness	Kurtosis
Mix1	-4.7e-005	0.007	0.5702	5.3239
Mix2	-3.98e-005	0.019	0.6317	5.9420

3.1 Time-Frequency (T–F) Masking

In this paper, the T-F masking separation is made use which is a powerful approach for the separation of underdetermined mixtures, especially for the separation of single-channel mixtures. The mask can be formed by comparing mixes spectrum with a threshold and thus forming a binary mask. Techniques based on time frequency masking use a time-frequency representation of the signal, taking profit from the disjointness provided by sparse transformations. Their aim is to identify the dominating source in each time-frequency unit, obtaining a mask. This mask dominates our desired source in the mix.The separated signals {y11....y1M},...,{yN1.....yNM} in frequency domain are constructed by time-frequency (T–F) masking

$$ykj(\iota, f) = Mk(\iota, f)Xj(\iota, f) \quad (9)$$

where, 0<M(ι,f)<1 is a mask specified for each separated signal yk and each time frequency slot(ι,f). For the design of masks, we rely on the sparseness property of source signals. A sparse source can be characterized by the fact that the source amplitude is close to zero most of the time. A time-frequency-domain speech source is a good example of a sparse source. Observation vectors X(ι,f) for all time-frequency slots are clustered into N classes C1,.....,CN each of which corresponds to a source signal .A vector X(ι,f) should belong to class Ck if the source sk is the most dominant in the observation X(ι,f) . We perform the clustering in a soft sense. A posterior probability, which represents how likely the vector x belongs to the kth class, is calculated in the "clustering". The T-F masks are specified by

$$Mk(\iota, f) = \{1, if P(Ck/x) > P(Ck'/x), k' = k \\ \{0, otherwise \quad (10)$$

In other words,the k^{th} mask Mk at a time-frequency slot (ι,f) is specified as 1 only if the k^{th} source is estimated as the most dominant source in the observation X at the T-F slot.To eliminate the effect of source amplitude we normalize them so that they have a unit norm. In the E-step, posterior probabilities are calculated and further modified in the m-step. EM algorithm estimates true probability distributions over both the direction from which sounds originate and the regions of the time-frequency plane associated with each sound source.

4. EVALUATIONS

4.1 Experimental Set Up

To verify the effectiveness of the proposed method,experiments with three sources which are mixed at two microphones at reverberation time of 150 ms and 250 ms at sampling rate of 16KHz is used. The arrangement taken should satisfy the undetermined condition that is number of sources should be more than the number of microphones.The experimental conditions are summarized in Table 4. The frame size depends on the sampling rate selected.

The separation performance is evaluated in terms of signal-to-distortion ratio (SDR). The Speech signals are noted to have more SDR value compared to songs or instrumental music and thus the separation is more effective for speech signals.The SDR measurements for various sources are summarized in Table 5 under reverberation time of 150ms and 250ms.

Table4. Experimental conditions

Number of microphones	M=2
Number of sources	N=3
Source signals	2 speech and 1 music
Reverberation time	150, 250 ms
STFT frame size	L=2024(16KHz)
Sampling rate	16KHz
STFT frame shift	512(16khz)

The SDR values indicate that separation results of speech mixtures are good compared to separation performance of music mixtures.The recorded mixtures provide greater SDR values than synthetic mixtures.The SDR equation is given:

$$SDR = 10log_{10}(Asignal/Anoise)^2 \quad (11)$$

Table.5 SDR Measurements

Sources	RT=150ms		RT=250ms	
	Mix 1	**Mix 2**	**Mix 1**	**Mix 2**
Female,Female ,Female	7.456	7.3789	6.6685	6.6159
Male, Male, Male	6.9795	6.9572	6.2541	6.2780
Music, Music, Music	2.9899	3.2245	2.4387	2.6495
Male, Female, Music	5.0321	4.7450	4.3466	4.1038
Female, Female, Male	6.6424	6.5916	5.9575	5.9054

The microphone observation sampled at a sampling frequency are converted into frequency-domain time-series signals by an STFT with appropriate sample frame and

sample shift.A number of mixes can be formed by same source combinations in different reverberation conditions. In this paper only experimental results with reverberation

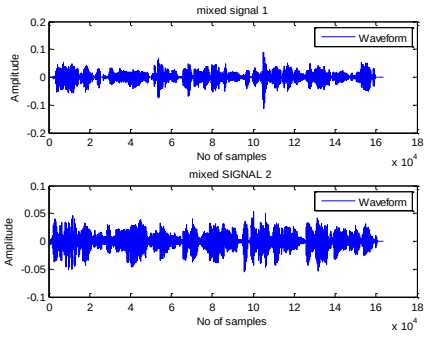

Fig.8.Plot of mixed signals

time of 150ms nad 250 ms are considered with male and female sources under different combinations.

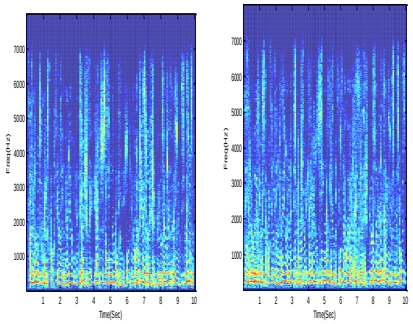

Fig.9.Spectrogram of mixed signals

The mixed signal is framed and converted to frequency domain by STFT. In Fig 9 and Fig 10 first frame of the two mixes are shown. On each frame the separation operation is carried out.

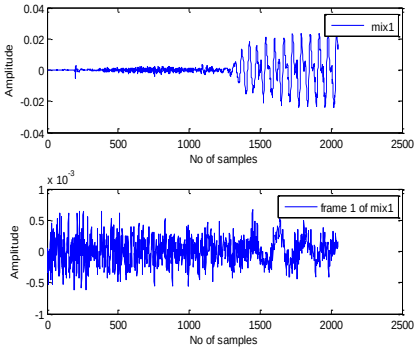

Fig.10.Plot of frame 1 of first mix

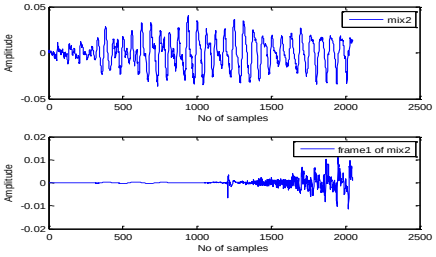

Fig.11. Plot of frame 1 of second mix

The mixed signals are framed and each is multiplied by a mask to get the original signal.

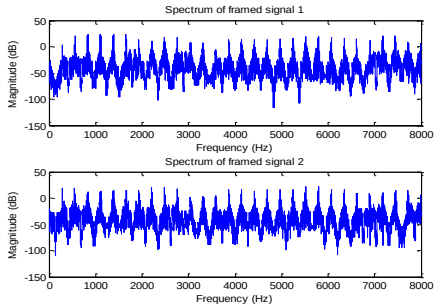

Fig.12. Spectrum plot of framed signals

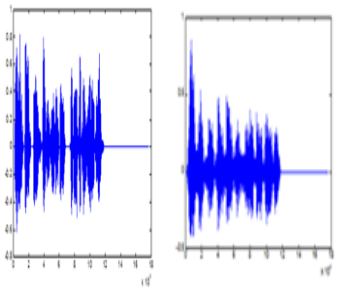

Fig.13.Plot of separated signals

The separation process gives the dominant source in the desired direction as output.

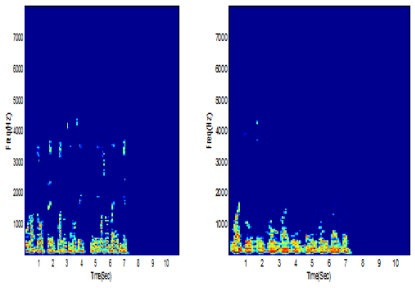

Fig.14.Spectrogram of separated signals

5. CONCLUSION

Thus the Blind Source Separation for various combinations of sources is attained using the two stage structure of the clustering part which considerably improves the separation performance. The experimental result indicates that in

evidence with the SDR values, separation results of speech mixtures(7.9 dB) are good compared to music mixtures(2.9 dB). The separation results for the benchmark data set andvarious recorded combination of mixtures are also performed under different reverberation conditions by T-F masking. In future, the independent component analysis (ICA) which effectively separates the music mixtures can also be implemented.

6.ACKNOWLEDGEMENT

I would like to thank my guide Mr.Sugumar D for his sincere support and being approachable and available whenever I needed any assistance. Above all I render my gratitude to the Almighty God who bestowed self confidence, ability and strength in time to complete this task.

7. REFERENCES

[1]. A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 3, pp. 550–563, Mar. 2010.

[2]. R. Olsson and L. Hansen, "Blind separation of more sources than sensors in convolutivemixtures," in Proc. ICAP'06, May 2006, vol. V, pp. 657–660.

[3] .O.Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process., vol. 52, no. 7, pp. 1830–1847, Jul. 2004

[4]. E.Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in Proc. ICA'09, 2009 [Online]. Available: http://sisec2008.wiki. irisa.fr/tiki-index.php.

[5].H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 5, pp. 1592–1604, Jul. 2007.

[6]. R. Mukai, S. Araki, H. Sawada, and S. Makino, "Evaluation of separation and dereverberation performance in frequency domain blind source separation," Acoust. Sci. Technol. vol. 25, no. 2, pp.119–126, 2004.

[7]. Netabayashi, Tomoyuki, "Robustness of the blind source separation with reference against uncertainties of the reference signals".

[8]. Blind Source Separation Of More Sources Than Mixtures Using Overcomplete Representations**,** Te-Won Lee, Member, IEEE, Michael S. Lewicki, Mark Girolami, Member, IEEE, and Terrence J. Sejnowski, Senior Member, ieee**,** IEEE signal processing letters, vol. 6, no. 4, april 1999 87.

[9]. **"**Comparison of Blind Source Separation Methods based on Iterative Algorithms J. Rinas, K.D. Kammeyer Department of Communications Engineering, Universitat¨ Bremen, frinas,kammeyerg@ant.uni-bremen.de

[10] ."Equivalence between Frequency Domain Blind Source Separation and Frequency Domain Adaptive Beamformers", Shoko Araki,Shoji Makino,Ryo Mukai,Hiroshi Saruwatari.NTT Communication Science Laboratories.

[11]. Equivalence between Frequency Domain Blind Source Separation and Frequency Domain Adaptive Beamformers, Shoko Araki,Shoji Makino,Ryo Mukai,Hiroshi Saruwatari.NTT Communication Science Laboratories.

[12].Ngoc Q. K. Duong, Emmanuel Vincent and Remi Gribonval "Under-determined reverberant audio source separation using a full-rank spatial covariance model" IEEE transactions on audio, speech, and language processing.