SINGLE MIXTURE AUDIO SOURCE SEPARATION USING KLD BASED CLUSTERING OF INDEPENDENT BASIS FUNCTIONS

[1]Md. Khademul Islam Molla, [1]Keikichi Hirose and [2]Nobuaki Minematsu

[1]Graduate School of Frontier Sciences, [2]Graduate School of Information Science and Technology
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
E-mail: {molla, hirose, mine}@gavo.t.u-tokyo.ac.jp

ABSTRACT

In this paper, we present a technique to separating the audio sources from a single mixture. The system is based on the extraction of independent basis function from the mixture spectrogram and grouping them to produce the source subspaces. Principal component analysis is used for dimension reduction and independent component analysis is employed here to make the basis functions independent from each other. Kullback-Leibler divergence (KLD) based information theoretic clustering algorithm is introduced in this work. The proposed algorithm is suitable for better grouping of the basis functions to separate the individual source. The satisfactory result of two-source mixture separation motivates to use this system for real world single mixture source separation.

1. INTRODUCTION

An effective system for separating the individual acoustic sources from their mixture would greatly facilitate many applications including automatic speech recognition (ASR), speaker verification, like automatic music transcription, broadcast news analysis and speaker separation in videoconference. Extracting multiple source signals from a single mixture is a challenging research field in the signal-processing arena.

The problem of identifying unknown sources in a mixture is called blind signal separation (BSS) and has found utility in many signal-processing applications [1, 2]. One of the different approaches to BSS is called Independent Component Analysis (ICA) [3]. ICA algorithms perform best when the number of observed signals is greater than or equal to the number of sources [1, 3, 4].

To overcome the restrictions, various extensions to the basic ICA have been proposed [1,2,5,6] in order to reduce the number of mixtures smaller than the number sources. Independent subspace analysis (ISA) method is used in [1, 4, 5] as the basic tool in audio source separation. In [7] and [8] the author proposed oscillatory correlation and pitch based technique respectively for monaural speech separation. The joint analysis of acoustic and modulation frequencies is applied as the key term in [9] for audio source separation from the mixture.

In this work, the time domain input mixture is projected on to the time-frequency (spectrogram) representation. The principal component analysis (PCA) is used to reduce the spectrogram dimension and ICA is applied to produce some independent basis functions. We have introduced and applied KLD based k-means clustering to group the independent basis functions into the source subspaces. Each subspace correspond individual source and the time domain source signal is derived by applying some inverse transformations. We have simulated our proposed method to separate the audio sources from the mixture speech with other signals and the output is with a satisfactory separation result.

Regarding the organization of this paper, the proposed separation algorithm is described in detail in section two, section three produces some experimental results. The discussion and some concluding remarks are presented in section four.

2. PROPOSED SEPARATION SYSTEM

An overall structure of the proposed method is shown in Figure 1. The following sub-sections describe the each component of the system structure. The source subspace decomposition operates on a one dimensional source mixture signal composed of N independent sources,

$$s(t)=\sum_{j=1}^{N}s_j(t) \tag{1}$$

The mixture signal is mapped to the spectral domain representation by Short Time Fourier Transform (STFT). From the complex valued spectrogram of n bins and m frames, the absolute value X and the phase information ϕ are separated. The phase information is re-used for the re-synthesis of the extracted sources.

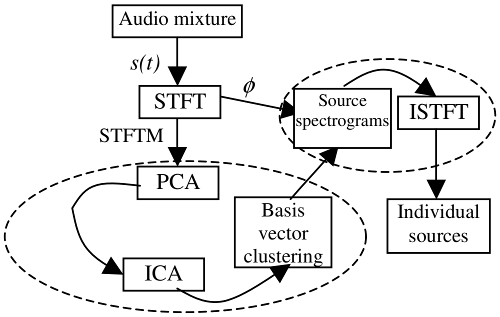

Figure 1: The overall system structure

The overall spectrogram *X* can be represented as the superposition of *N* unknown independent spectrograms as:

$$X = \sum_{i=1}^{N} x_i \tag{2}$$

x_i is also uniquely represented as the outer product of a invariant frequency basis function f_i, and a corresponding amplitude envelope function a_i which describe the variation of the frequency basis function over time [5].

$$x_i = f_i a_i \tag{3}$$

So by summing all component spectrogram:

$$X = \sum_{i=1}^{N} f_i a_i \tag{4}$$

This assumption corresponds that the frequency basis functions are stationary within the spectrogram i.e. no pitch change occur. Casey, Westner in [2] suggested to decompose the source signal into some block of stationary spectral distribution.

The independent basis functions represent the features of the independent sources, and each source is composed of some independent basis functions as in Equation (3). After producing the independent basis functions, clustering method is applied to make the subsets of basis function corresponding to each source. In this paper we suggest an information theoretic approach of k-means clustering algorithm to group the basis functions.

2.1 Dimension Reduction

The number of rows and columns of the magnitude spectrogram X are greater than the number of sources in a mixture. PCA is used here to reduce the dimension of X. The Singular Value Decomposition (SVD) is a well-defined generalization of the PCA [10]. A singular value decomposition of X is any factorization of the form:

$$X_{m\times n}=U_{m\times m}S_{m\times n}V_{n\times n}{}^{T} \tag{5}$$

where U and V are column wise orthogonal matrices, each column is termed as principal component (PC). S is a diagonal matrix of singular values with components $\sigma_{i.}$ U contains the principal components of X based on frequency bins, while V contains the principal components based on time frames. Some of the principal components are selected (from U or V) as the basis functions. The set of required basis functions can be taken from U or V. If it is taken from U, after applying ICA the basis will be independent in terms of frequency bins otherwise the basis will be independent of time frames.

The SVD orders the basis vectors by the size of their singular values $\sigma_{i.}$ The ith singular value represents the amount of information contained in the ith principal components. The required number of PCs to separate the sources can be estimated by the inequality:

$$\frac{1}{\sum_{i=1}^{n}\sigma_i}\sum_{i=1}^{p}\sigma_i \geq \varphi \tag{6}$$

where σ_i is the singular value of the ith PC, φ is the threshold (amount of information in %) and p is the number of required basis function.

The amount of required information is also very much application specific. Our experiment suggests that 45-55% (φ=0.45 to 0.55) can successfully recover the sources from the mixture of two signals.

2.2 Independent Basis Functions

The basis vectors obtained by PCA are only uncorrelated but not statistically independent. To derive the independent basis vectors a further procedure called ICA must be carried out. The ICA model [3] expresses the observation signal x as the product of mixing matrix A and vectors of statistically independent signals,

$$x = As \qquad (7.1)$$

where A is (pseudo-) invertible mixing matrix and s is random signal vector. The ICA algorithm estimates W so that the output signals u are as statistically independent as possible.

$$u = Wx = WAs \qquad (7.2)$$

In this model, x corresponds the basis functions obtained from PCA and u is the collection of independent basis functions. There exist a number of ICA algorithms to be used depending on the nature of experimental data [3, 11]. JadeICA algorithm [11] is applied here to determine the demixing matrix W.

Once the independent basis functions have been obtained, the corresponding amplitude envelopes or frequency basis functions can be obtained by projecting the magnitude spectrogram on to the plane of the independent basis functions. Then both types of basis are grouped into the desired number of sources.

2.3 Independent Basis Functions

Usually more basis vectors than sources are selected during the dimension reduction using PCA to keep the maximally informative source subspace. We have introduced Kullback-Leibler divergence (KLD) based k-means clustering algorithm for partitioning the independent basis functions into source subspaces. Symmetric KLD measures the relative entropy between two probability mass functions p(x) and q(x) over the random variable X as:

$$KLD(p,q) = \frac{1}{2}\{\sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} + \sum_{x\in X} q(x)\log\frac{q(x)}{p(x)}\} \qquad (8)$$

The KLD(.) always takes a non-negative value for p≠q, and zero for p=q. KLD is used here to measure the information theoretic distance during k-means clustering. Each basis function is normalized and transformed to its corresponding probability mass function. The proposed algorithm is given Figure 2.

Traditional k-means algorithm performs clustering based on Euclidean distance. It implies that the data clusters are ball-shaped and very much affected by the scaling factor. The variation in amplitudes with the same shape of independent basis vectors is supposed to be happened because ICA algorithm does not guarantee to recover the signals at their original amplitude [3]. Whereas, KLD based measure compares the relative information as a clustering tool and hence it is unaffected by scaling. It is more effective for independent basis vector clustering to produce the source subspaces.

```
Initialize k cluster center weights $w_1, w_2 \ldots w_k$
Repeat steps 1 to 3 until convergence is reached
   For iteration $t$:
1.  Select a normalized basis function $v$
2.  (a) Calculate the distances $d_j$ of $v$ from $w_j$
          $d_j=KLd(v,w_j), j=1, 2, 3\ldots.k$
    (b) Identify the center $i$ closest to $v$
          $i=arg\ min\{d_j\}, j=1, 2,3\ldots k$
3.  (a) Update the weight of the $i^{th}$ center
          $w_i(t+1)=w_i(t)+\eta(t)(v-w_i(t))$
    (b) Update the learning rate factor $\eta$
          $\eta(t+1) = \frac{\eta(t)}{1 + 0.0005\ t^{0.02}}$
    (c) Calculate the change $C$ of weights by
          $C = \frac{\| \vec{w}(t+1) - \vec{w}(t) \|_2}{\| \vec{w}(t+1) \|_2}$
4. Check: If $C<\varepsilon$ then stop, else repeat.
```

Figure 2: KLD based k-means clustering algorithm

While the grouping of the basis functions (frequency and time) into the number of sources the magnitude spectrogram of individual source is constructed by using Equation (3). The rest is only to re-synthesis the sources.

2.4 Re-synthesis of the sources

The magnitude spectrogram x_i of individual source is already derived as the outer product of i^{th} group of basis functions (f_i and a_i). The source spectrogram S_i is obtained by simply inserting the phase matrix ϕ of the original source spectrogram to x_i as:

$$S_i = x_i \cdot e^{j[\phi]} \qquad (9)$$

The corresponding time domain source signals are produced by applying the Inverse STFT on the source spectrograms.

3. EXPERIMENTAL RESULTS

We have used some mixture of two audio streams to test the efficiency of our proposed separation algorithm. The individual test stream is a mixture of male speech and other sound like female speech, violin, telephone ring, jazz music, white noise etc. All mixtures are with 11025Hz sampling rate and 8-bit amplitude resolution. We have applied the separation algorithm on the audio segments with no large variance of pitch and then concatenate the extracted source signals to produce the separation over the entire stream.

Average value of the short-term signal to mixture (called average SMR) is used here for quantifying the separation efficiency. The SMR can be defined as:

$$SMR\ (t) = 10 \log 10 \left(\frac{\sum_{i=0}^{w-1} s^2(t+i)}{\sum_{i=0}^{w-1} m^2(t+i)} \right) \qquad (10)$$

where s and m are signal and the mixture with other signal respectively, w is window length and it is 10 ms here. The experimental separation efficiency (SMR) of some mixtures are presented in Table 1.

Table 1: The experimental separation results of our proposed algorithm. Sig1 is male speech signal and sig1 represents other signal as indicated.

Mixtures of male speech+	SMR (original)		SMR (extracted)	
	sig1	sig2	sig1	sig2
Flute sound	-8.23	-12.33	-13.63	-16.76
Tel ring	-9.42	-15.31	-12.12	-18.21
Jazz music	-4.22	-6.13	-14.83	-15.82
Femalespeech	-3.11	-4.12	-11.21	-18.18

The separation efficiency measured by SMR is actually relative measure. It indicates the distances of the original and separated signals from the mixture. Smaller the difference between original and separated SMR indicates better separation. The separation efficiency is better for speech and telephone ring mixture than any other one. It is argued that the separation performance is higher for those sources with some differences in spectral distribution.

4. DISCUSSION AND CONCLUSIONS

A data-adaptive single mixture audio source separation method is presented in this paper. We have considered the mixture of two stationary audio sources. There are many researches on audio source separation using microphone array and the research trend is to reduce the number of microphone. There are some preliminary researches on single mixture separation [2, 5] and most of the works are training based. The algorithm proposed here is able to separate the audio sources from their single mixture without any prior knowledge about the sources. Hence it enhances the state of the art in audio source separation arena.

An entropy-based approach of k-means clustering is introduced here to grouping the independent basis functions to build the source subspaces. It is not affected the amplitude variation that is usually happens during the application of ICA to basis vectors. It successfully groups the independent basis vectors into a given number of source subspaces.

To enhance the robustness, to derive the strong criteria to select the number of basis function required, automatic detection of the number of sources in the mixture are the future plan of this work.

REFERENCES

[1] Christian Uhle, Christian Dittmar, Thomas Sporer, "Extraction of Drum Tracks from Polyphonic Music using Independent Subspace Analysis", ICA2003, Nara, Japan, April 2003.

[2] Casey M.A., A. Westner, "Separation of Mixed Audio Sources by Independent Subspace Analysis", International Computer Music Conference, 2000.

[3] A. Hyvärinen and E. Oja, "Independent Component Analysis: Algorithms and Applications", Neural Networks, 13(4-5): 411-430, 2000.

[4] Orife Riddim, "A rhythm analysis and decomposition tool based on independent subspace analysis", Masters thesis, Darthmouth College, 2001.

[5] Derry FitzGerald, Eugene Coyle, Bob Lawlor, "Sub-band Independent Subspace Analysis for Drum Transcription", International Conference on Digital Audio Effects, Germany, 2002.

[6] Tuomas Virtanen, "Sound Source Separation Using Sparse Coding with Temporal Continuity Objective", ICMC, 2003.

[7] DeLiang L. Wang, Guy J. Brown, "Separation of Speech form Interfering Sounds Based on Oscillatory Correlation", IEEE Transaction of Neural Network Vol. 10, No. 3, May 1999

[8] Guoning Hu, DeLiang Wang, "Monaural Speech Separation", Proceedings of Neural Information Processing System (NIPS'02), 2003.

[9] Les Atlas, "Modulation Spectral Transforms-Application to Speech Separation and Modification", Technical Report of IEICE, SP2003-51, 2003

[10] Casey M.A., "Auditory Group Theory: with application to statistical basis methods for structured audio", PhD thesis, MIT Media Laboratory, 1998

[11] J.F. Cardoso, A. Souloumiac, "Blind beamforming for nonGaussian signals", IEEE Proceedings, Vol. 140, no. 6, pp. 362-370, 1993.