2005 RISP International Workshop on Nonlinear
Circuit and Signal Processing (NCSP'05)
Hawaii, USA, Mar. 4-6, 2005.

SINGLE MIXTURE SEPARATION OF SPEECH AND INTERFERING AUDIO SIGNALS USING SUBSPACE DECOMPOSITION

Md. Khademul Islam Molla[†], Keikichi Hirose[‡] and Nobuaki Minematsu[†]

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

Abstract

This paper presents a method of separating speech and interference signal from their single mixture. The system is based on deriving some independent basis vectors from the mixture spectrogram and clustering them to produce the individual source subspaces. Principal component analysis (PCA) is used to derive some basis vectors reducing the dimension of the mixture spectrogram and independent component analysis (ICA) is used to make the basis vectors independent in their own domain. The independent basis vectors are then grouped into two sets (speech and interfering audio) by employing Kullback-Leibler divergence (KLd) based k-means clustering. Each group of basis vectors is used to decompose the mixture spectrogram into the individual source spectrograms and the time domain source signals are re-synthesized by applying some inverse transformations. The experimental results are noticeable in separating speech and its interfering audio signals.

1. Introduction

The segregation of speech signal from the interfering audio signals from their mixture(s) has been taken attention regarding many signal processing applications including automatic speech recognition (ASR), and music transcription. The performance of ASR degrades quickly in presence of noise and other interfering signals [1]. To mitigate the effect of noise, signal sources are filtered out by spectral subtraction based method [2], computational auditory scene analysis (CASA) system [3, 4], and so on. The primary grouping cues used in most CASA systems is F0 – this works well only for parts of the speech signal that contain voiced components. The spectral subtraction method also requires some spectral information about the source signals. In many realistic applications, the performance of those approaches is inadequate. Moreover, both speech and interfering signals may be subject to consider for in some applications. The separation algorithm proposed here is fully data adaptive and can separate speech and interfering audio signals without any prior knowledge about the sources.

Independent component analysis (ICA) is one of the approaches in data adaptive blind source separation (BSS) method and ICA algorithm performs best when the number of mixture signals is greater than or equal to the number of sources [5, 6, 7]. Although some recent over complete (mixtures < sources) representation relax this assumption [7], separating sources from only one mixture remains problematic. The oscillatory correlation is used in [3] as the basic cue of separation. Hu and Wang [4] introduced a speech segregation system from single mixture using pitch tracking and observing common amplitude modulation scenario. Their system can only separate the voiced speech from the mixture. Roweis [8] proposed learning based statistical pattern recognition process to separate the sources from single mixture. Casey [9] proposed a method to represent auditory group theory with statistical basis functions. In single mixture situation, the spectral distribution of the candidate signals is the principal cue of source separation.

In this paper, some basis vectors are derived from the single mixture spectrogram using PCA to perform the audio sources separation. Two types of basis vectors: time basis and frequency basis can be derived from the spectrogram. The suitable one is selected for better separation depending on the energy distribution characteristics of the mixture spectrogram. Applying ICA, the independent basis vectors are grouped into the number of sources using KLd based clustering. Each group is used to derive the source subspaces and some inverse transformation is applied to synthesis time domain source signals. Regarding the organization of this paper, we have described the separation algorithm in detail in section two, some experimental results are presented in section three, and section four contains discussion and some concluding remarks

2. Proposed Separation Algorithm

The block diagram of the overall separation algorithm is shown in Figure 1. The source subspace decomposition operates on the audio mixture signal *s(t)* composed of *N* independent sources. Here *N=2* corresponds speech and only one interfering signal e.g. single noise source. The mixture signal is then projected onto the time-frequency plane *S(f,t)* using short time Fourier transform (STFT). We can easily compute the magnitude and phase information as, $X(f,t) = |S(f,t)|$ and $\phi(f,t) = \arg[S(f,t)]$. *X(f,t)* ≡ *X* is referred as the magnitude spectrogram (*MS*) which is the data space to derive two sets of independent basis vectors. Each column represents the spectrum of the windowed time frame, and the rows represent the amplitude variation of the

spectrums over time. The phase matrix $\phi(f,t)$ is saved to be used in re-synthesizing of the extracted sources.

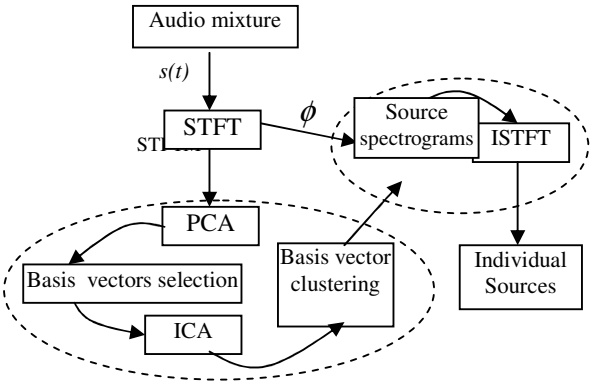

Figure 1: The block diagram of the separation algorithm

The overall magnitude spectrogram X can be represented as the superposition of N independent source spectrograms as: $X = \sum_{i=1}^{N} x_i$ .The x_i is also uniquely represented as the outer product of frequency invariant basis vector $B_i = [b_1^{(i)}, b_2^{(i)} .....b_n^{(i)}]$ and corresponding amplitude envelope (basis vector invariant of time frame) $A_i = [a_1^{(i)}, a_2^{(i)} .....a_n^{(i)}]^T$ (n is the number of basis components to represent x_i) which describes the magnitude variation of the frequency basis vectors over time [10] as $x_i = B_i A_i$.

The number (n) of basis components depends on the spectral characteristics of the individual source. The principal objective is to derive two sets of basis vectors from the mixture *MS* yielding the separation of speech and interfering source. The frequency basis vector (B_i) is assumed as stationary over all the time frames and time basis vector (A_i) is also static for all frequency bins. This assumption implies that the mixture signal is framed into some segments to hold the above consideration.

2.2. Deriving Basis Vectors

Principal component analysis (*PCA*) is used to reduce the dimension of *MS* producing a number of basis vectors. The PCA linearly transforms a set of correlated variables into a number of uncorrelated variables called principal components (*PC*) [7]. The Singular Value Decomposition (SVD) is a well-defined generalization of the PCA [9] and become an important tool in statistical data and signal processing. A singular value decomposition of an $l \times k$ matrix X is any factorization of the form:

$$X = UDV^T \qquad (1)$$

where $U_{l \times l}$ and $V_{k \times k}$ are orthogonal (with orthogonal column) matrix and D is an $n \times k$ diagonal matrix of singular values. If X corresponds the *MS* the columns of U contains the PCs of X based on frequency bins, while V contains the PCs based on time frames.

Each singular value represents the amount of information contained by corresponding PC. A reduced set of PCs are selected as the basis vectors to derive the source subspace from the mixture *MS*. The cumulative sum of the singular values carry important cue to select the number of PCs as basis vectors with maximally informative subspaces [11]. Considering the experimental results, the number of basis vectors corresponding to 50-60% of the total amount of information can successfully separate the sources from the mixture of two signals.

2.3. Selection of Basis Vectors

A reduced set of PCs can be selected from U or V to make them independent by applying ICA. If taken from U the basis vector after ICA will be independent of frequency channel and for the case of V it will be independent of time frames. It is a crucial decision to select basis vectors from U or V and this selection affects the separation result. In [6, 9], the source stream is divided into a number of segments with constant spectra, and frequency independent basis vectors (from U) are used. In [10], it is suggested to divide the mixture stream into some blocks with almost constant pitch and proposed to pick the basis vectors any one (U or V).

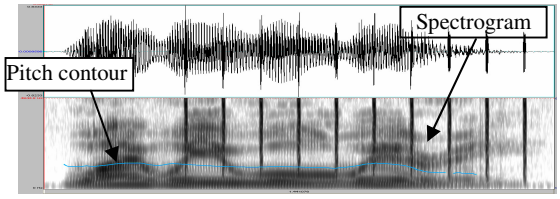

Figure 2: Spectrogram of a mixture of speech bip-bip sound with the pitch contour.

Now consider the situation as shown in Figure 2; the spectrogram of a mixture of speech and some bip-bip sounds. The pitch is almost constant over the entire spectrogram. But practically the sources are not separated with independent basis vectors selected from U. On the other hand, if we like to break the spectrogram as suggested into some segment with constant spectra, it creates a huge number of source segments (at least each segment for each bip sound) and not easy to recombine after separation. The time independent basis vectors (selected from V) are suitable to separate the sources considering the whole spectrogram as single segment.

We propose the criteria: (i) to select the basis vectors from when the energy distribution is more regular and continuous along the frequency channels; (ii) to select from V with the energy distribution more regular, dense and continuous along the time frames. The proposed contrast functions used to

determine the time or frequency basis vectors are described here. The contrast function along the frame direction is derived as:

$$Cr_t(s_t, c_t, d_t) = \sum_t s_t c_t d_t \qquad (2)$$

where s_t is the frame similarity, c_t is the energy continuity and d_t is the denseness (reciprocal of sparseness) of energy along the frames. In the similar way the contrast function along frequency channels is calculated in the same manner as:

$$Cr_f(s_f, c_f, d_f) = \sum_f s_f c_f d_f \qquad (3)$$

If the value of Cr_f is greater than the value of Cr_t, the basis vectors selected from *U* are passed to *ICA* algorithm to make them independent. Otherwise the vectors are selected from V. We will discuss about the components and their underlying significances in the upcoming full-length paper. The quantitative selection criteria of some mixture streams are shown in Table 1. Observing the Table 1, it is clear that mixture stream s4 (mixture of speech and bip-bip sound) is a candidate of using time basis vectors (from *V*) to derive the independent source subspaces. This quantitative criterion has a great role to improve the robustness of single mixture source separation.

Table 1: Basis vector selection criteria of some mixture files

Audio	Mixture of	Cr_t	Cr_f
s1	Speech & Tel ring	0.769	2.202
s2	Speech & jazz music	0.565	1.667
s3	Male & female speech	1.121	1.563
s4	Speech & bip-bip sound	6.375	1.385
s5	Speech & flute sound	1.437	3.810

2.4. Constructing Independent Source Subspaces

The basis vectors obtained by *PCA* (selected from *U* or *V* depending on the value of the proposed contrast function) are only uncorrelated but not statistically independent. To derive the independent basis vectors a further procedure called *ICA* must be carried out. JadeICA algorithm [12] is used here to estimate the independent version of the reduced set of *PC*s (basis vectors). It should be noted that *ICA* is applied on only one type of frequency (from *U*) or time (from *V*) basis vectors producing the set $B = [b_1, b_2 ..... b_M]$ or $A = [a_1, a_2 ..... a_M]^T$ respectively with *M* is the reduced number of vectors using *PCA*. The corresponding *A* or *B* are obtained by projecting the *X* on *B* or *A* as: $A=B^TX$ or $B=XA^T$ respectively.

The components of each basis vector *B* and *A* are grouped into B_i and A_i subsets respectively. For a two source mixture problem $i=1,2$ i.e. two subsets (B_1, B_2) and (A_1, A_2). Then the magnitude spectrogram of individual source (speech and interfering source) is constructed as $x_1=B_1A_1$ and $x_2=B_2A_2$.

We have introduced a Kullback-Laibler divergence (KLd) based *k*-means clustering algorithm to bunch the independent basis vectors into *k* groups (here $k=2$). Symmetric KLd measures the relative entropy between two probability mass functions $p(x)$ and $q(x)$ over a random variable *X* as:

$$KLd(p,q) = \frac{1}{2}\{\sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} + \sum_{x\in X} q(x)\log\frac{q(x)}{p(x)}\} \qquad (4)$$

Each basis vector is normalized and transformed to its corresponding probability mass function. Then *KLd* is used for distance measure between two basis vectors during k-means clustering whereas traditional *k*-means measures the Euclidean distance. *KLd* being information theoretic measure performs better.

The source spectrogram s_i is calculated by inserting the phase information of the original mixture spectrogram as:

$$s_i(f,t) = x_i(f,t)\cdot e^{j[\phi(f,t)]} \qquad (5)$$

The corresponding time domain source signals are produced by applying inverse STFT to each of the source spectrogram.

3. Experimental Results

We have simulated our system to separate the sources from the mixture of two audio streams (speech and other sounds). The individual test stream is a mixture of speech and other sounds as indicated in Table 1. All mixtures are with 16kHz sampling rate and 16-bit amplitude resolution. As the parameters of STFT, 512 point FFT with 30ms width and 20ms overlapping Hamming window are used. 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.

The average value of the running short-term relative energy between original and separated signal is used here for quantifying the separation efficiency. It is termed here as original to separated signal ratio (*OSSR*) and mathematically defined as:

$$OSSR = \frac{1}{T}\sum_{t=1}^{T} \log 10\left(\frac{\sum_{i=1}^{w} s_o^2(t+i)}{\sum_{i=1}^{w} s_r^2(t+i)}\right) \qquad (6)$$

where s_o and s_r are the original and separated signal respectively, *T* is the total time length, *w* is a 10ms square window. This *OSSR* calculates the relative short-term energy level between those two signals and used to measure the

difference between two signals in terms of short-term energy level. If the two signals are exactly similar, the *OSSR* produces 0 value and any other value (positive or negative) is a measure of their dissimilarity. Table 2 shows the average *OSSR* of each signal for every mixture. Smaller deviation of *OSSR* from 0 indicates the higher degree of separation. Based on quantitative evaluation, it is observed that the separation efficiency of the proposed algorithm is noticeable in separating two sources from single mixture.

Table 2: The Experimental results of our proposed algorithm. Sig1 is male speech and sig2 is any other sound as in Table 1.

Mixtures	***OSSR of Sig1***	***OSSR of Sig2***
s1	-0.2203	-0.0620
s2	-0.3609	0.1328
s3	0.4520	0.1770
s4	0.2763	-0.0920
s5	-0.2802	0.1212

4. Discussion and Conclusions

A data-adaptive single mixture source separation method is proposed without any prior knowledge about the sources. Some existing single mixture source separation methods [2, 4, 8] have employed priori knowledge about the sources and emphasized on segregating only speech signal from interfering source. In [6, 10], a part of the mixture e.g. components from polyphonic music are extracted with some priori training. This algorithm is able to separate and re-synthesis both speech and interference without any information regarding the source characteristics.

We have proposed a selection criteria of the basis vector direction i.e. frequency or time which basis vector is the candidate to be independent in their own domain for better source separation. This selection criterion enhances the robustness of the single mixture source separation method. Some exiting separation algorithms as in [9, 10] of single channel separation suggest to divide the mixture signal into some segments with static pitch or stationary spectra and to use the frequency independent basis vector. But they are unable to separate the sources when the energy distribution in the spectrogram segment of static pitch is more regular along time frame (as shown in Figure 2). Our proposed contrast function of basis vector selection is more applicable in such situation.

An entropy-based approach of k-means clustering is introduced here to grouping the independent basis vectors 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. We need some more post-processing for enhancement of the audio quality of the separated sources. The automatic detection of the number of sources in a given mixture, their proper separation, and to employ the separation algorithm in practical uses are the main concern as our future works.

References

[1] Y. Gong, "Speech recognition in noisy environments: A survey", Speech Communication, vol. 16, pp. 261-291, 1995

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Transaction on ASSP, vol. 27, pp. 113-120, 1979

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

[4] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation", IEEE Transaction of Neural Networks, in press, 2004.

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

[6] C. Uhle, C. Dittmar, and T. Sporer, "Extraction of Drum Tracks from Polyphonic Music using Independent Subspace Analysis", ICA2003, Nara, Japan, April 2003.

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

[8] S. T. Roweis, "One Microphone Source Separation", NIPS, pp. 793-799, 2000.

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

[10] D. FitzGerald, E. Coyle, and B. Lawlor, "Sub-band Independent Subspace Analysis for Drum Transcription", International Conference on Digital Audio Effects, Germany, 2002.

[11] J. F. C. L. Cadima and I. T. Jolliffe, "Variable selection and the interpretation of principal subspaces" *J. Agric. Biol. Environ. Stat.*, 6:62-79, 2001.

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