THE ROBUSTNESS AND APPLICABILITY OF AUDIO SOURCE SEPARATION FROM SINGLE MIXTURES

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

[1] Graduate School of Information Sciences and Technology

[2] Graduate School of Engineering

The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

*Address of Corresponding Author: molla@gavo.t.u-tokyo.ac.jp

ABSTRACT: The separation of audio sources from their single mixture is a great challenge in signal processing research. Many single mixture source separation techniques have been proposed in the past 20 years but unfortunately the results are not pleasing enough for practical applications. In this tutorial-review paper, single-channel audio source separation techniques are divided into three broad categories: separation by auditory scene analysis (ASA), training based separation and blind source separation (BSS). Each of the categories is briefly described to contrast their methodological differences. This study focuses on the limitations and robustness under adverse acoustic environment of the seveal categories. We compare the success and usability of the different techniques in real world applications.

1. INTRODUCTION

The technical challenge of blind separation of audio sources is actively pursued in both engineering and applied mathematical disciplines. Single mixture (monaural) blind source separation addresses the most challenging case. It has many applications in signal processing including automatic speech recognition and music transcription. Monaural separation of acoustical sources refers to an algorithm that separates the components from a mixture of acoustical signals [1]. It is especially useful in circumstances where multiple sources are closely spaced and therefore where methods based on spatial localisation fail.

The single mixture situation often happens during taping of speakers' utterances in a public space with a single microphone. The problem is that the information of the audio sources (locations, the acoustics of the surrounding place, energy ratios, phase-delays, etc.) is merged in a single channel. All information related to the target source is mixed up with the information of the interfering sources. Researches on single mixture source separation are aiming on several directions. It is difficult to categorise the techniques explicitly to illustrate their comparative studies. A well established method is commonly known as auditory scene analysis (ASA) [2, 3, 4, 5]. Some techniques (i.e., training based method) use *a priori* knowledge about the sources of some specific types of mixtures [6, 7]. The statistical signal processing approach, such as independent component analysis (ICA) [8], is also the subject of increasing research. In this paper, we discuss three broad categories of single mixture audio source separation methods. We have the robustness and applicability of the different categories in a practical scenario.

2. CLASSIFICATION

Despite considerable research on single mixture audio source separation, the success has been limited. In the following sub-sections, three categories of research regarding single channel source separation are discussed: (i) computational auditory scene analysis (ii) training based separation, and (iii) separation by independent subspace analysis (ISA). Each category has its own success, applicability and failure.

2.1 Auditory Scene Analysis (ASA)

Auditory scene analysis (ASA) is the process by which the human auditory system organises sound into perceptually meaningful elements. . Computational ASA (CASA) is a machine learning system that aims to separate mixtures of sound sources in a way similar to that used by human listeners. It is closely related to audio source separation. CASA differs from the field of blind source separation in that, like the human auditory system, it uses no more than two microphone recordings of an acoustic environment. The block diagram of CASA system is shown in Figure 1.

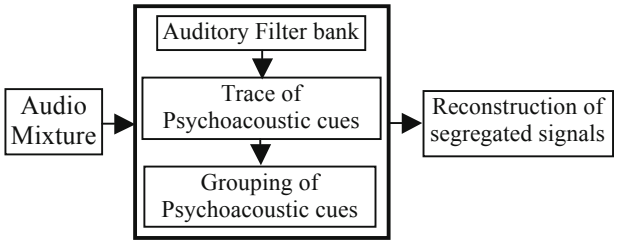

Figure 1: A block diagram of Computational Auditory Scene Analysis (CASA).

In traditional CASA, the goal is to extract multiple sound components from a mixture. The output sound can then be analysed independently to compute their features. The separated sounds should be similar to the input sound in some perceptual ways. Scheirer [2] proposed an ASA-based method of sound understanding without separation similar to the ability of the human listener. ASA understands a real environment using acoustic events. The human auditory system can easily solve some ASA problems in a multi-source audio environment. But,

in solving the problem of ASA using acoustic signals received from the same environment, a unique solution cannot be derived without constraints on acoustic sources and the real environment. The sequential steps of the auditory segregation model are shown in Figure 1.

2.1.1 *Auditory model of mixed audio signal*

The audio mixture is first analysed into a time-frequency representation by an auditory filter-bank approximating the function of the cochlea. Different types of filter-bank, for example, constant-Q filter-banks [3] and gammatone filter-banks [4] have been suggested. Bregman [5] reported that, for ASA, the human auditory system uses four psychoacoustically heuristic regularities related to acoustic cues namely, (*i*) common onset and offset, (*ii*) gradualness of change, (*iii*) harmonicity, and (*iv*) changes in the acoustic event. Tracing and grouping of such cues in the time-frequency representation are the most important steps in source segregation using CASA.

2.1.2 *Auditory grouping*

The mixed audio signal is first decomposed into a collection of acoustic components. The auditory system appears to construct the subsets of analysed components using some organisational principles [4]. The organisation explicitly represents the form of groups of auditory primitives which can reasonably be assumed to belong together.

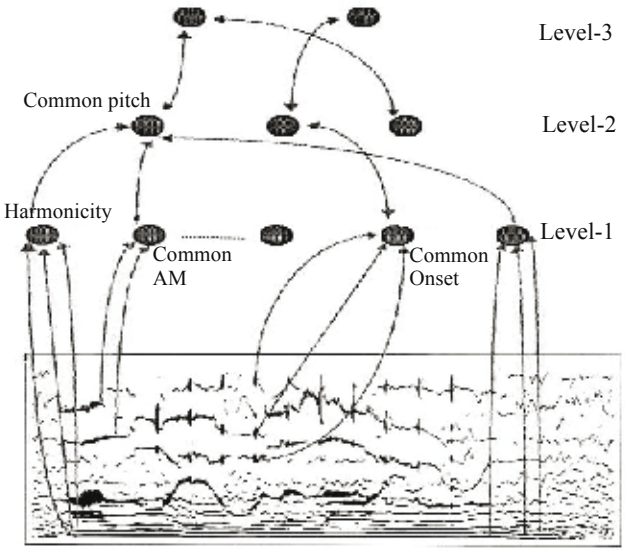

Figure 2: A hierarchical view of auditory scene analysis.

When those components are considered together, the grouping is performed from a complete and self-consistent explanation of the listener's acoustic environment. The hierarchical model of auditory grouping process is shown in Figure 2.

If the initial decomposition has resulted in a set, $C=\{c_1, c_2,\ldots\ldots\ldots,c_n\}$, of components, the grouping task might be expressed as one which demands that a collection $G=\{g_1, g_2, \ldots\ldots, g_m\}$ of groups be discovered, such that,

$$\forall(g \in G) \cdot g \subseteq C \tag{1}$$

This approach is to view grouping as a process, which builds a hierarchical model on the set *C*, as illustrated in Figure 2. The resulting representation is a graph in which lowest level (Level-1) nodes correspond to auditory primitives, whilst intermediate level (Level-2) nodes represent the groups of objects discovered at the preceding level (Level-1). The upper level (Level-3) would correspond to more complete explanations of each acoustic source – auditory streams.

The grouping process consists of more or less direct implementation of the cues known to us from psychoacoustics [3]. Cues such as harmonicity, common onset, common amplitude modulation (AM), and energy continuity [5] are used to group the acoustical elements. For each rule, the grouping method takes a single element as a seed. The grouping process returns zero or more groups to grow up from the seed. When the grouping is properly completed, the next step is to partition the time-frequency representation of the mixture into the number of sources. Ideally the number of groups with all the acoustic cues is equal to the number of sources forming the mixture. Most of the time, some clean-up or post-processing is required for proper grouping and segregation of the sources.

2.2 Training-Based Separation

Considering the complexity of the blind source separation, some researchers have proposed training based segregation of sources from a single mixture. Some use time domain methods [8, 9] and others time-frequency based algorithms [10]. Training based separation requires either strong assumptions about the nature of the sources, substantial *a priori* information, or a combination of both.

2.2.1 *Time domain method*

Most time domain techniques are based on splitting the whole signal space into several disjoint and orthogonal subspaces that suppress overlaps. The criteria employed by the earlier time domain methods mostly involve second order statistics (SOS). Those methods perform well with input signals well-suited to the AR (autoregressive) model. Moreover, the use of SOS restricts the separable cases to orthogonal subspaces [11]. The recent approach [8, 9] based on exploiting *a priori* sets of time domain basis functions learned by independent component analysis (ICA) uses higher-order statistics (HOS) resolving the problems with SOS. For better understanding, this paper presents the time domain separation method described in [8, 9].

To formulate the problem, the observed signal y^t is assumed as the summation of *P* independent source signals,

$$y^t = \lambda_1 x_1^t + \lambda_2 x_2^t + ... + \lambda_P x_P^t \tag{2}$$

where $t \in [1, T]$, x^t_i is the t^{th} sampled value of the i^{th} source signal, and λ_i is the gain of each source which is fixed over time. The goal is to recover all x^t_i given only a single sensor input y^t. To represent the generative model, continuous samples of length *N* with $N<<T$ are chopped out of a source; the subsequent segment is denoted as an *N*-dimensional column vector x_i^t, expressed as

$$x_i^t = \sum_m^M a_{im} s_{im}^t = A_i s_i^t \quad (3),$$

where M is the number of basis functions (a_{im}) and s^t_{im} is the coefficients of the vector x^t_i. It is assumed that $M = N$ with A as a full rank, reversible matrix. The ICA learning algorithm is used to determine $W_i = A_i^{-1}$ such that $s^t_i = W_i x^t_i$. Figure 3 shows the generative model of time domain signal decomposition as the weighted sums of the basis functions. The algorithm first involves the ICA-based learning of time domain basis functions of the sources that are the subjects to be separated. This corresponds to the prior information necessary to separate successfully the signals [8]. Considering P=2 and hence, $\lambda_1+\lambda_2=1$, only λ_1 is estimated. The proposed method is tested with four different types of sounds: rock music, jazz, male and female speech with reasonable success rates.

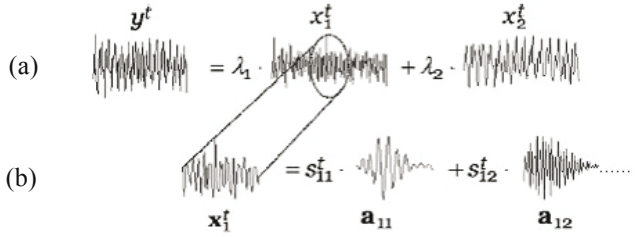

Figure 3: Generative model of the observed mixture and the source signals; (a) mixture as the weighted sum of two source signals, and (b)individual source generated by weighted (s^t_{im}) linear superposition of basis functions (a_{im}).

2.2.2 *Method in time-frequency domain*

To tackle the underdetermined problem, single mixture source separation is often implemented using short-time spectra, i.e. a time-frequency representation of the mixed signal. In such a representation, we may consider sound as comprising a collection of localized time/frequency components. The distribution of these basic sound elements on the time-frequency plane will effectively be the spectrogram of the analysed sound. Using this representation, we can employ a wealth of probabilistic analysis techniques directly on the spectral distributions without having to worry about enforcing non-negativity and also providing a clear way to incorporate these techniques in learning framework [11].

The basis decomposition with the time-frequency method is of the form $y_i=w_i.S$, where S is the input magnitude spectrogram in the form of matrix and w_i is a matrix containing the weight corresponding to the i^{th} source, and y_i is a magnitude spectrogram of the i^{th} source signal. If we wish to separate the i^{th} source component, we can then modulate the phase of the original mixture spectrogram with y_i and invert the frequency transform yielding the separated source. The method for computing the weight factors can be varied. As an example, the monaural source separation method by Roweis [10] is illustrated here. In [10], a filtering technique is presented to estimate time-varying masking filter that localise sound streams in a spectro-temporal region. The sources are supposedly disjoint in the spectrogram and the estimated mask exclusively divides the mixed streams completely. If $w_k(t)$ is a set of masking (weighting) functions, a source signal $y(t)$ can be recovered by modulating the corresponding sub-band signals $X_k(t)$ as [10]:

$$y(t) = \sum_{k=1}^{K} w_k(t) X_k(t) \quad (4),$$

where K is the number of sub-band signals. The masking method is performed on the original spectrogram as shown in Figure 4. The goal of the source separation depends on the construction of the masking signals $w_k(t)$, i.e. to group together regions of the spectrogram that belong to the same auditory sources.

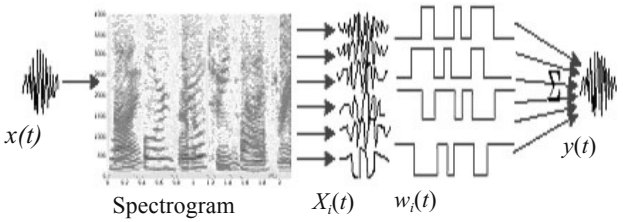

Figure 4: Masking approach of source separation

Roweis [10] uses simple factorial hidden Markov models (HMM) system which learns from the spectrogram of a single speaker to generate the masking function. The approach first trains speaker-dependent HMM on isolated data from single speakers. Then, to separate a new single recording which is a mixture of known speakers, these pre-trained models are combined into a factorial HMM (FHMM) architecture. The FHMM consists of two or more underlying Markov chains which evolve independently and the sources are separated simultaneously. The results of separating a simple two-speaker (male & female) mixture are presented in [10]. Training based source separation, being based on prior knowledge about the sources, is not able to adapt many robust situations which are common in real-world application such as robust speech recognition, signal de-noising etc.

2.3 Separation by ISA

Presently, several audio source demixing researchers [12, 13, 14, 15] proposed independent subspace analysis (ISA) method as the tool of separation. ISA aims to derive some independent basis vectors from the single mixture spectrogram (time-frequency plane). Two types of basis vectors, temporal basis and spectral basis, can be derived from the spectrogram. The more suitable one for source separation is selected based on the energy distribution in the spectrogram [13]. Applying ICA, the independent basis vectors are then grouped together to derive the subspaces corresponding to each source. Kullback-Laibler divergence-based clustering algorithm is introduced in [13] to group the independent basis vectors into the number of sources. The time-frequency representation can also be performed on a Hilbert spectrum [15], a newly developed method for analysing non-linear and non-stationary signals.

2.3.1 *Basic separation model using ISA*

The block diagram of the overall separation algorithm is shown in Figure 5. The source subspace decomposition operates on the audio mixture signal $s(t)$ composed of N

independent sources. The mixture signal is then projected onto the time-frequency plane $\boldsymbol{S}(n,k)$ using short time Fourier transform (STFT) [13] or a Hilbert spectrum [15]. One can easily separate magnitude $X(n,k)$ and phase $\phi(n,k)$ information from $S(n,k)$.

The overall magnitude spectrogram X can be represented as the superposition of N independent source spectrograms as:

$$X = \sum_{i=1}^{N} \boldsymbol{x}_i \tag{5}$$

x_i is also uniquely represented as the outer product of an independent spectral basis vector F_i, and corresponding amplitude envelope A_i (temporal basis vector) as:

$$x_i = F_i A_i^T \tag{6}$$

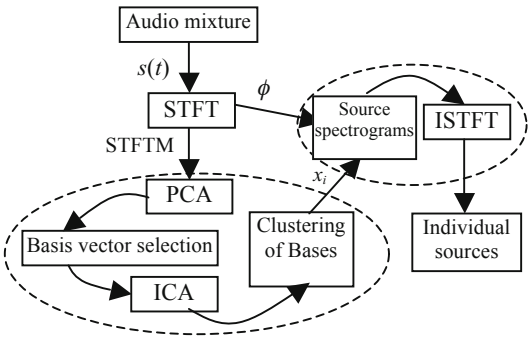

Figure 5: The block diagram of the separation algorithm using ISA.

Now the object is to derive N sets of F_i and A_i from X using the sequential application of principal component analysis (PCA), independent component analysis (ICA) and clustering method. Each set of basis vectors corresponds to features of the independent sources.

2.3.2 *Constructing independent subspaces*

Usually, the number of rows and columns of X is greater than the number of spectral/temporal basis vectors required for subspace decomposition. The dimension of the overall magnitude spectrogram X is first reduced by principal component analysis (PCA) [12, 14].

The basis vectors obtained by PCA are only uncorrelated but not statistically independent. To derive the independent basis vectors, a further procedure called ICA is be carried out. JadeICA algorithm [15] is used here to obtain the independent basis vectors (F or A based on selection in PCA). Once the spectral or temporal independent basis vectors have been obtained, the corresponding amplitude envelopes A or frequency basis F respectively can be obtained by projecting X on to the independent one (F or A). The basis vectors are then grouped (group F and A into F_i and A_i subsets respectively) into the number of sources (for two sources i=1, 2 i.e. two subsets of F and A). In [13], a Kullback-Laibler divergence (KLd) based k-means clustering algorithm has been proposed for the grouping process. After properly grouping F and A, the individual source subspaces (source spectrograms) are obtained by Eq. (6). The time domain source signal is constructed by simply appending the phase matrix $\phi(n,k)$ and applying the inverse STFT (ISTFT). The separation results of two sources from single mixture using ISA [13] are illustrated in Figure 6.

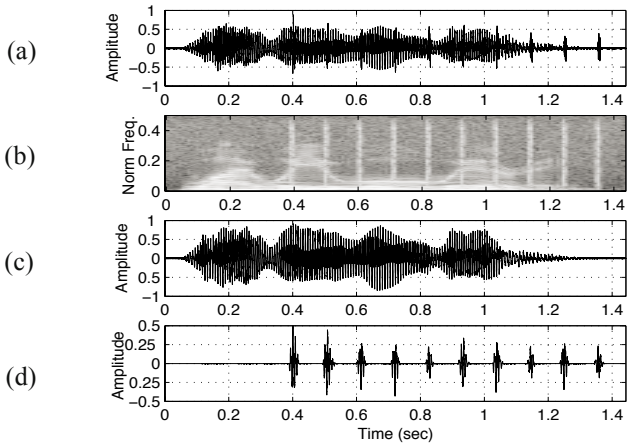

Figure 6: Separation by ISA [13]; (a) mixture signal of speech and bubble noise, (b) spectrogram of the mixed signal, (c) separated speech signal and (d) separated bubble noise.

We [13] have presented the simulation results to separate the sources from the mixture of two audio streams (speech and other sounds). Using this method, it is not possible to separate two similar sources such as two male speakers. In this case, the proposed model is likely to produce the same type basis vectors, indicating similar source features.

3. COMPARATIVE DISCUSSION

Single-mixture audio source separation techniques do not work well when the spectra of the component sources overlapp. None of existing methods of single mixture source separation is capable of separating two or more speech signals coming from moving male speakers recorded through a single microphone. It is difficult to compare the methods discussed above. Table 1 provides a comparison of the above-mentioned methods regarding their robustness and usability in practical applications.

REFERENCES

1. G. Cauwenberghs, "Monaural Separation of Independent Acoustical Components", *Proc. IEEE International Symposium on Circuits and Systems (ISCAS'99)*, Orlando FL, 1999.
2. E. D. Scheirer, "Sound Scene Segmentation by Dynamic Detection of Correlogram Comodulation", *MIT Media Laboratory Technical Report* No. 491, April 1999.
3. D. PW. Ellis, " A Computer Implementation of Psychoacoustic Grouping Rules'', *Proc. 12th Int. Conf. on Pattern Recognition*, 1994.
4. M. Cooke, "Modelling Auditory Processing and Organisation'' *Cambridge University Press, Cambridge,* 1993.
5. A. S. Bregman, "Auditory Scene Analysis: hearing in complex environments'', *pp. 10-36, Oxford University Press, New York,* 1993.

Table 1: A comparison of Single-Channel audio source separation methods

Separation Method	Robustness	Applicability
Auditory scene analysis	a. There are many ambiguous situations for sources with similar spectra. b. Performance also depends on the time-frequency representation of the mixture signal.	a. More suitable to segregate only one source from the mixture. b. Applicable for vocal extraction from music, speech segregation from noise etc.
Training based separation	a. Being training-based, it focuses on a specific region of source separation problem. b. Unable to separate sources with similar spectral characteristics. c. Implemented to date only for the mixture of two sources.	a. Requires a *priori* knowledge about the sources. b. Real world applications include robust speech recognition, signal de-noising of known signal-noise type.
Separation by ISA	a. Performs better when there are some differences spectra of the component sources. b. Performance depends on the time-frequency analysis and segmentation of the mixture signal.	a. Separation is performed without any prior knowledge about the sources. b. Suitable when the sources are independent and linearly mixed.

6. G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation", *IEEE Trans. Neural Networks*, 15, 1135-1150, 2004.
7. M. J. Reyes-Gomez, D. Ellis and N. Jojic, "Multiband Audio Modeling for Single Channel Acoustic Source Separation", *Proc. ICASSP-04,* May 2004.
8. G. -J. Jang, T. –W. Lee and Y. –H. Oh, "Single-Channel Signal Separation Using Time-Domain Basis Functions", *IEEE Signal Processing Letters*, Vol. 10, No. 6, June 2003.
9. G. -J. Jang and T. –W. Lee, "A probabilistic approach to single channel source separation" *Advances in Neural Information Processing Systems 15, MIT Press, Cambridge*, 2003.
10. S. T. Roweis, "One Microphone Source Separation", *NIPS*, pp. 793-799, 2000.
11. S. Makino, T.-W. Lee and H. Sawada, "Blind Speech Separation", *Springer*, 2007.
12. M. A. Casey, and A. Westner, "Separation of Mixed Audio Sources by Independent Subspace Analysis", *International Computer Music Conference*, 2000.
13. M. K. I. Molla, K. Hirose, N. Minematsu, "Separation of speech and interfering audio signal from single mixture by subspace decomposition", *Journal of Signal Processing*, Vol. 9, No. 6, pp: 487-495 (2005).
14. C. Uhle, C. Dittmar and T. Sporer, "Extraction of Drum Tracks from Polyphonic Music using Independent Subspace Analysis", *Proc. of ICA2003*, Nara, Japan, 2003.
15. M. K. I. Molla and K. Hirose, "Single-mixture audio source separation by subspace decomposition of Hilbert spectrum", *IEEE Transaction on Audio, Speech and Language Processing,* Vol. 15, No. 3, pp. 893-900, 2007.