Chapter 3
Cocktail Party Problem:
Source Separation Issues and Computational Methods

Tariqullah Jan
University of Surrey, UK

Wenwu Wang
University of Surrey, UK

ABSTRACT

Cocktail party problem is a classical scientific problem that has been studied for decades. Humans have remarkable skills in segregating target speech from a complex auditory mixture obtained in a cocktail party environment. Computational modeling for such a mechanism is however extremely challenging. This chapter presents an overview of several recent techniques for the source separation issues associated with this problem, including independent component analysis/blind source separation, computational auditory scene analysis, model-based approaches, non-negative matrix factorization and sparse coding. As an example, a multistage approach for source separation is included. The application areas of cocktail party processing are explored. Potential future research directions are also discussed.

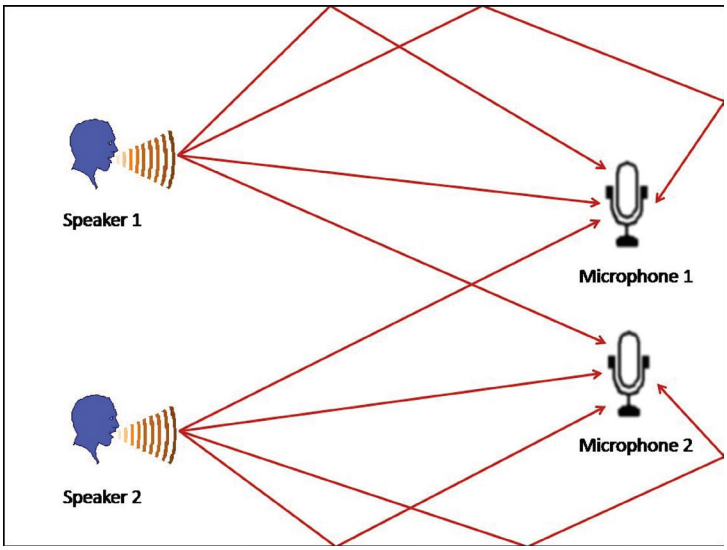
INTRODUCTION

The concept of the cocktail party problem (CPP) was coined by Cherry (1953). It was proposed to address the phenomenon associated with human auditory system that, in a cocktail party environment, humans have the ability to focus their listening attention on a single speaker when multiple conversations and background interferences and noise are presented simultaneously. Many researchers and scientists from a variety of research areas attempt to tackle this problem (Bregman, 1990; Arons, 1992; Yost, 1997; Feng et al., 2000; Bronkhorst, 2000). Despite of all these works, the CPP remains an open problem and demands further research effort. Figure 1 illustrates the cocktail party effect using a simplified scenario with two simultaneous conversations in the room environment.

As the solution to the CPP offers many practical applications, engineers and scientists have spent their efforts in understanding the mechanism of the human auditory system, and hoping to design a machine which can work similarly to the human

DOI: 10.4018/978-1-61520-919-4.ch003

Figure 1. A simplified scenario of the cocktail party problem with two speakers and two listeners (microphones)



auditory system. However, there are no machines produced so far that can perform as humans in a real cocktail party environment. Studies on the human auditory system could help understand the cocktail party phenomenon, and offer hopes of designing a machine that could approach a normal human's listening ability.

It has been observed that people with the perceptive hearing loss suffer from insufficient speech intelligibility (Kocinski, 2008). It is difficult for them to pick up the target speech, in particular, when there exist some interfering sounds nearby. However, amplification of the signal is not sufficient to increase the intelligibility of the target speech as all the signals (both target and interference) are amplified. For this application scenario, it is highly desirable to produce a machine that can offer clean target speech to these hearing impaired people.

Scientists have attempted to analyze and simplify the complicated CPP problem, see, for example, a recent overview in (Haykin, 2005). A variety of methods have been proposed for this problem. For example, computational auditory scene analyses (CASA) approach attempts to transform the human auditory system into mathematical modeling using computational means (Wang & Brown, 2006; Wang, 2005). Blind source separation (BSS) is also used by many people to address this problem (Wang et al., 2005; Araki et al., 2003; Olsson et al., 2006; Makino et al., 2005). BSS approaches are based on the independent component analysis (ICA) technique assuming that the source signals coming from different speakers are statistically independent (Hyvarinen et al., 2001; Lee, 1998). Non-negative matrix factorization (NMF) and its extension non-negative tensor factorization (NTF) have also been applied to speech and music separation problems (Smaragdis, 2004, Virtanen, 2007; Schmidt & Olsson, 2006, Schmidt & Laurberg, 2008, Wang, 2009). Another interesting approach is the sparse representation of the sources in which the source signals are assumed to be sparse and hence only one of the source signals in the mixture is active while others are relatively insignificant for a given time instant (Pearlmutter et al., 2004; Bofill et al., 2001; Zibulevsky & Pearlmutter, 2001). Some model based approaches have also been employed to address this problem (Todros et al., 2004; Radfar

et al., 2007). The following sections provide a detailed review of these techniques for addressing the cocktail party problem, in particular, for audio source separation which is a key issue for creating an artificial cocktail party machine.

BACKGROUND FOR AUDIO SOURCES

Audio sources are usually classified as speech, music or natural sounds. Each of the categories has its own specific characteristics which can be exploited during its processing. Speech sounds are basically composed of discrete phonetic units called phonemes (O'Shaughnessy, 2000; Deng & O'Shaughnessy, 2003). Due to the co-articulation of successive phonemes, each signal that corresponds to a specific phoneme exhibits time varying properties. The resultant signal is composed of periodic harmonic pulses which are produced due to the periodic vibration of the vocal folds, a noise part which is generated because of the air passing via lips and teeth, or a transient part due to the release of pressure behind lips or teeth. Harmonics within the generated signal has periodic frequency components which are the multiples of a fundamental frequency component. In real speech signals the fundamental frequency component of the periodic phonemes varies due to the articulation, but typically for male speech is 140 Hz, and 200 Hz for female speech with variation of 40 Hz for each.

Music sources (Hall, 2001) generally constitute of sequences of notes or tones produced by musical instruments, singers and synthetic instruments. Each note is composed of a signal which further can be made of a periodic part containing harmonic sinusoids produced by blowing into pipe or bowing a string, or a transient part generated due to hitting a drum or plucking a string, or a wideband noise produced by blowing the wind instruments. For example, in western music the periodic frequencies of the notes generated typically remain constant or varying slowly. Musical instruments usually produce musical phrases which are composed of successive notes without any silence between the notes. Unlike monophonic music, polyphonic sounds are composed of several simultaneous notes that are generated by multiple musical instruments.

The third source comes from the environment, called natural sounds (Gygi et al., 2004). Their characteristic varies depending on the origin of the natural sound. Similar to the speech and music signals it can also be classified as periodic, transient and noise. For example, a car horn produces the natural periodic sound signal, a hammer thrashing the hardwood generates the transient signal and raining results in a wideband noise signal. The discrete structure of natural sound is simpler as compared with the organization of notes and phonemes. In this chapter, we will mainly focus on the first type of the audio source signal i.e. speech signals. The methods for the CPP discussed in this chapter are mainly applied in context of the speech signals.

COMPUTATIONAL AUDITORY SCENE ANALYSIS

The ear is mainly composed of three parts: the outer ear, the middle ear and the inner ear. The outer ear constitutes of a flap of tissue which is visible and called pinna, and the auditory canal (Mango, 1991). The combination of pinna and auditory canal helps in sound source localization. Sound moving through the auditory canal results in the vibration of eardrum within the middle ear. The middle ear transmits these vibrations to the inner ear. The middle ear, which is composed of three small bones i.e., the malleus, incus, and strapes, plays an important role in the transmission of vibrations. The middle ear is an impedance matching device between the air and fluid-filled inner ear. Inside the inner ear there is an organ called cochlea containing fluid. The vibrations

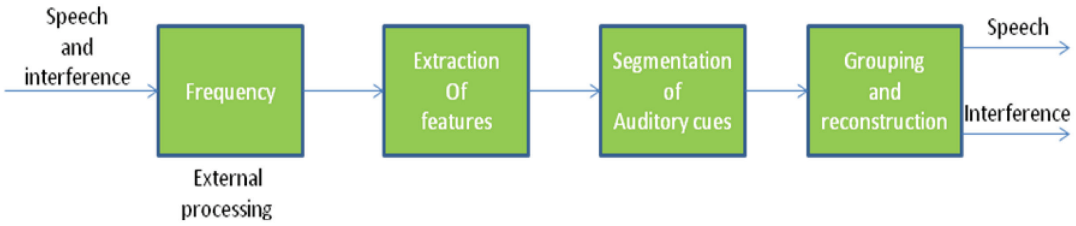
transferred into the inner part of the ear press the cochlear fluid and hence stimulate the hair cells in the cochlea. Hair cells have a role of converting physical vibrations into a set of nerve responses (i.e. electrical signals), and they are frequency-selective, which means that different regions of the cochlea, more precisely different areas of basilar membrane, response to different frequencies (hair cells at the end near the oval window correspond to high frequency up to 20 kHz, and those in a narrow part to low frequency). Therefore the cochlea performs a kind of spectral analysis and can be modeled as a bank of band-pass filters (Auditory scene analysis: listening to several things at once, 2007). Electrical signals from the cochlea are transferred as neural impulses along the auditory nerve towards the brain. The central parts of the auditory system are relatively complex and less understood in comparison to periphery.

CASA is the study of auditory scene analysis (ASA) by computational means (Wang & Brown, 2006). ASA is the process by which the human auditory system organizes sound into perceptually meaningful elements. The concept of ASA was coined by Bregman (1990). The human auditory system is complicated and constitutes of two ears and auditory routes (Haykin et al., 2005). Specifically it is a refined system which has the ability to distinguish the frequency components coming from different sources and also can find the exact location for the source signals. This ability of the human auditory system is very unique because of the fact that the frequency component arrangement inside the signal and the combination of signals is very perplexing. Generally speaking, the human auditory system performs sound localization and recognition in order to pick up the target signal from the cocktail party environment. In literatures we can find different approaches for the localization of sound signal, for example (Blauert, 1983; Yost, 2000). Time difference, level difference and spectral difference are the important acoustic cues used for the localization of sound sources. The recognition can be well explained from the work presented by Bregman (1990). According to his analysis recognition can be done in two major steps called sound segregation and sound determination. Segregation of sound sources can be achieved using feature selection and feature grouping. Feature selection consists of some very important features like pitch. Feature grouping basically combines the incoming sound components in such a way that a stream of similar components corresponding to a single sound source is grouped together. Sound determination is then performed to identify the elements within the sound stream rather than just segregation.

Hence CASA systems are machine listening systems that aim to separate mixtures of sound sources in the way that the human auditory system does. The fundamental steps required in order to segregate the speech signal by CASA systems are: First, to analyze the signals in such a way that the interfering speech signals can be neglected. In the second step, a recognition process is involved where the speech signals mixed in a stream are analyzed according to their statistical property that is important for recognizing the target signal. The last step called synthesis involves reorganizing the target signals from the separated sound stream. CASA approaches have been employed to investigate the cocktail party problem (Wang et al., 2006; Wang, 2005; Cooke et al., 2001, Cooke, 2002). The architecture of a typical CASA system is shown in Figure 2.

In general, there are two types of approaches for the separation of the target signal in the cocktail party environment in the context of CASA. The first one is called "signal-driven" approach which is used for the segregation of the auditory scene into the different components belonging to the different sound streams (Bregman, 1990). The second one called "knowledge-driven" approach uses the prior knowledge of the unknown speech sources, so that the target signal can be separated from the interference. In 1994, Brown and Cooke investigated some of the key issues related to the early CASA methods (Brown & Cooke, 1994).

Figure 2. Schematic diagram of a typical CASA system



Specifically they avoid the assumptions made about the type and number of sources. They proposed to model the human auditory system into separate parts. The key parts are ear filtering, cochlear filtering and central processing (combination of different auditory maps which shows onset, offset, periodicities and frequency transitions). Wang and Brown (1999) extended the work of Brown and Cooke by replacing the central processing with a double layer oscillator network and applied simple computational methods for auditory feature extraction.

A technique called ideal binary masking (IBM) has been recently used in CASA to segregate the target signal from the interference (Wang et al., 2006). Consider a microphone signal recorded in a cocktail party problem: $x(t) = s_1(t) + s_2(t)$, where $s_1(t)$ is the target speech signal and $s_2(t)$ is the interference speech signal and t is the discrete time instant. Denote X, S_1 and S_2 as the time-frequency (T-F) representation of $x(t)$, $s_1(t)$ and $s_2(t)$, obtained from some T-F transformation respectively. Then the ideal binary mask (IBM) for $s_1(t)$ with respect to $s_2(t)$, is defined as follows,

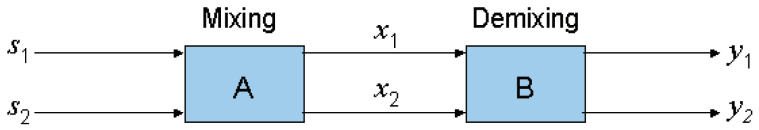
$$M_1(t,f) = \begin{cases} 1 & \text{if } |S_1(t,f)| > |S_2(t,f)| \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The target speech $s_1(t)$ can then be extracted by applying the IBM to X, followed by an inverse T-F transform. The decision is binary, and hence the intelligibility of the segregated speech signal is high. But on the other hand the resultant mask M_1 entirely depends on the availability of the target and interference speech signals. In practice, the target and interference signals are usually unknown, and the mask has to be estimated from the mixtures.

BLIND SOURCE SEPARATION

Another technique to address the cocktail party problem is BSS, where the mixing process is usually described as a linear convolutive model and convolutive ICA algorithms can then be applied to segregate the source signals from their mixtures assuming the sources are statistically independent (Araki et al., 2003; Olsson & Hansen, 2006; Makino et al., 2005; Mitianondis & Davies, 2002; Nickel & Iyer, 2006; Pedersen et al., 2008). BSS is an approach used for the estimation of the source signals having only the information of the mixed signals observed at each input channel, without prior information about sources and the mixing channels. Its potential applications include speech segregation in cocktail party environment, teleconferences and hearing aids. In such applications, the mixture signals are reverberant, due to the surface reflections of the rooms. ICA is a major statistical tool for the BSS problem, for which the statistical independence between the sources is assumed (Hyvarinen et al., 2001; Lee, 1998). The mathematical model (Ainhoren, 2008) used to describe the ICA is given as,

Figure 3. Schematic diagram for a typical BSS system with two sources and two mixtures. Unknown source signals: s, observed signals: x, estimated signals: y



$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) + ... + a_{1N}s_N(t)$$
$$\vdots$$
$$\vdots$$
$$x_M(t) = a_{M1}s_1(t) + a_{M2}s_2(t) + ... + a_{MN}s_N(t) \quad (2)$$

where $s_1(t),....,s_N(t)$ representing unknown source signals in the cocktail party environment, $x_1(t),....,x_M(t)$ denote the mixture signals (e.g. microphone recordings). If the coefficients $a_{ij}(i = 1,\ldots,M$ and $j = 1,\ldots,N)$ are scalars, the resultant mixtures are referred to as instantaneous mixtures, and if they are filters, the mixtures are referred to as convolutive mixtures. If $N = M$, i.e., the number of sources equals to the number of mixtures, it is called exactly determined BSS problem. If $N > M$, it is the under-determined case, and $N < M$ the over-determined BSS problem. A schematic diagram of a typical two input two output BSS system is given in Figure 3, in which A represents the unknown mixing system and B is the demixing system used for the estimation of the unknown source signals.

The BSS approach using ICA can be applied either in the time domain (Pedersen at al., 2008; Cichocki & Amari, 2002) or in the frequency domain (Wang et al., 2005; Araki et al., 2003; Olsson & Hansen, 2006; Makino et al., 2005) or their hybrid (Lee et al., 1997; Lambert & Bell, 1997), assuming that the source signals are statistically independent. The time-domain approaches attempt to extend the instantaneous ICA model to the convolutive case. They can achieve good separation performance once the algorithms converge, as the independence of segregated signals is measured accurately (Makino et al., 2005). However the computational cost for the estimation of the filter coefficients in the convolutive operation can be very demanding, especially when dealing with reverberant mixtures using long time delay filters (Amari et al., 1997; Matsuoka & Nakashima, 2001; Buchner et al., 2004; Douglas & Sun, 2002; Douglas et al., 2005).

To improve the computational efficiency, the frequency domain BSS approaches transform the mixtures into the frequency domain, and then apply an instantaneous but complex ICA algorithm to each frequency bin (Wang et al., 2005; Araki et al., 2003; Parra & Spence, 2000; Schobben & Sommen, 2002; Sawada et al., 2003; Mukai et al., 2004). As a result, many complex valued and instantaneous ICA algorithms that have already been developed can be directly applied to the frequency domain BSS. However an important issue associated with this approach is the permutation problem, i.e., the permutation in each frequency bin may not be consistent with each other so that the separated speech signal in the time domain contains the frequency components from the other sources. Different methods have been developed to solve this problem. By reducing the length of the filter in the time domain (Buchner et al., 2004; Parra & Spence, 2000) the permutation problem can be overcome to some extent. Source localization approach has also been employed to mitigate the permutation inconsistency (Soon et al., 1993; Sawada et al., 2004). Another technique for the alignment of the permutations across the frequency bands is based on correlation between separated

source components at each frequency bin using the envelope similarity between the neighboring frequencies (Murata et al., 2001).

The third approach is the combination of both time and frequency domain approaches. In some methods (Lee et al., 1997; Back & Tosi, 1994), the coefficients of the FIR filter are updated in the frequency domain and the non-linear functions are employed in the time domain for evaluating the independence of the source signals. Hence no permutation problem exists any more, as the independence of the source signals is evaluated in the time domain. Nevertheless, the limitation of this hybrid approach is the frequent switch between two different domains at each step and thereby consuming extra time on these inverse transformation operations.

The separation performance of many developed algorithms is however still limited, and leaves a large room for improvement. This is especially true when dealing with reverberant and noisy mixtures. For example in the frequency-domain BSS framework, if the frame length of the DFT is long and the number of samples in each frequency bin is small, the independence assumption may not be satisfied. Similarly, if the short length DFT frame is used, the long reverberations cannot be covered and hence the segregation performance is limited (Araki et al., 2003).

Apart from the above discussed methods, some authors consider the assumption of W-disjoint orthogonality for speech signals in order to separate the source signals from the observe data. For example in (Jourjine et al., 2000), for a given windowing function $W(t)$, two sources, $s_i(t)$ and $s_j(t)$ are called W-disjoint orthogonal if the supports of the short-time Fourier Transform of $s_i(t)$ and $s_j(t)$ are disjoint (Jourjine et al., 2000). The windowed Fourier Transform of $s_i(t)$ is defined as,

$$F^w\left[s_i\right](\tau, w) = \int_{-\infty}^{\infty} W(t-\tau)s_i(t)e^{-iwt}dt \qquad (3)$$

which can be denoted as $s_i^w(\tau, w)$. The W-disjoint orthogonality assumption can be expressed as below (Jourjine et al., 2000).

$$s_i^w(\tau, w) * s_j^w(\tau, w) = 0, \forall i \neq j, \forall w, \tau \qquad (4)$$

This equation implies that either of the sources is zero for any w and τ as long as two sources do not come from the same source. If $w(t) = 1$, then $s_i^w(\tau, w)$ can be interpreted as the Fourier Transform of $s_i(t)$, which can be referred to as $s_i(w)$. Therefore, W-disjoint orthogonality can be written as,

$$s_i(w) * s_j(w) = 0, \forall i \neq j, \forall w \qquad (5)$$

which represents the property of disjoint orthogonality (Jourjine et al., 2000).

Another challenging problem is to separate moving sources rather than stationary in a cocktail party environment. A recent work by (Naqvi et al., 2009) is devoted to the blind separation of moving sources. Here a multimodal approach is proposed for the segregation of moving speech sources. The key issue in blind estimation of moving sources is the time varying nature of the mixing and unmixing filters, which is hard to track in real world. In this work the authors applied the visual modality for the separation of moving sources as well as stationary sources. The 3-D tracker based on particle filtering is used to detect the movement of the sources. This method performs well for the blind separation of moving sources in a low reverberant environment.

So far, two important techniques for the CPP were discussed in detail. It is interesting to make a comparison between these two techniques. In the case of BSS, the unknown sources are assumed to be statistically independent. However, no such assumption is required for CASA. On the other hand, the IBM technique used in the CASA

domain needs to estimate the binary mask from the target and interference signals which should be obtained from the mixture in practice. Another difference is in the way how the echoes within the mixture are dealt with by these two techniques. In BSS algorithms (Wang et al., 2005; Araki et al., 2003; Olsson & Hansen, 2006; Makino et al., 2005), such a situation is modeled as a convolutive process. On the other hand CASA approaches deal with echoes based on some intrinsic properties of audio signals, such as, pitch, which are usually preserved (with distortions) under reverberant conditions. However, the human auditory system has a remarkable ability of concentrating on one speaker by ignoring others in a cocktail party environment. Some of the CASA approaches (Wang & Brown, 1999) work in a similar manner i.e. extracting a target signal by treating other signals as a background sound. In contrast, BSS approaches attempt to separate every source signal simultaneously from the mixture. Motivated by the complementary advantages of the CASA and BSS approaches, we have developed a multistage approach in (Jan et al., 2009, 2010) where a convolutive BSS algorithm is combined with the IBM technique followed by cepstral smoothing. The details of this method (Jan et al., 2009) will be discussed later in this chapter as an example.

MODEL BASED APPROACHES

Another method to address the cocktail party problem is based on the statistical modeling of signals and the parameters of the model are estimated from the training data. Some model based approaches have been used for the blind separation of speech signals e.g., (Todros & Tabrikian, 2004; Radfar & Dansereau, 2007; Ichir & Djafari, 2006; Radfar et al., 2006). In (Todros & Tabrikian, 2004) Gaussian mixture model (GMM) which is widely used for the modeling of the highly complex probability density functions (pdf), is employed for the modeling of the joint pdf of the sources by exploiting the non-gaussianity and/or non-stationarity of the sources and hence the statistical properties of the sources can vary from sample to sample.

In (Radfar & Dansereau, 2007) the model-based approach is used for single channel speech separation. The authors considered the problem as speech enhancement problem in which both the target and interference signals are non-stationary sources with same characteristics in terms of pdf. Firstly, in the training phase, the patterns of the sources are obtained using the Gaussian composite source modeling. Then the patterns representing the same sources are selected. Finally, the estimation of the sources can be achieved using these selected patterns. Alternatively, a filter can be built on the basis of these patterns and then applied to the observed signals in order to estimate the sources.

Source separation in the wavelet domain by model-based approaches has been considered in (Ichir & Djafari, 2006). This method consists of a Bayesian estimation framework for the BSS problem where different models for the wavelet coefficients have been presented. However there are some limitations with the model based approach. The trained model can only be used for the segregation process of the speech signals with the same probability distribution, i.e., the pdf of the trained model must be similar to that of the observation data. In addition, the model based algorithms can perform well only for a limited number of speech signals.

NON NEGATIVE MATRIX/ TENSOR FACTORIZATION

NMF was proposed by Lee & Seung in 1999. Using the constraint of non-negativity, NMF decomposes a non-negative matrix V into the product of two non-negative matrices W and H, given as:

$$V_{mxn} = W_{mxr} H_{rxn} \tag{6}$$

where $(n + m)\ r < mn$. Unlike other matrix factorizations, NMF allows only additive operations i.e. no subtractions (Lee & Seung, 1999, Lee & Seung, 2001, Laurberg et al., 2008). As NMF does not depend on the mutual statistical independence of the source components, it has a potential to segregate the correlated sources. NMF has been applied to a variety of signals including image, speech or music audio. In (Cichocki et al., 2006) the authors attempted to separate the general form of signals from the observe data i.e. both positive and negative signals using the constraints of sparsity and smoothness. For machine audition of audio scenes, NMF has also found some applications. For example, it has been applied to music transcription (Smaragdis & Brown, 2003, Wang et al, 2006) and audio source separation (Smaragdis, 2004, Smaragdis, 2007, Wang & Plumbley, 2005, Parry & Essa, 2007, FitzGerald et al, 2005, FitzGerald et al, 2006, Morup et al, 2007, Schmidt & Morup, 2006, Wang, 2007, Virtanen, 2007, Wang et al, 2008, Wang et al, 2009). In these applications, the audio data are usually transformed to non-negative parameters, such as spectrogram, which are then used as the input to the algorithms. The application of the NMF technique to the CPP problem is still an emerging area which attracts increasing interests in the research community. For an overview of recent progress on NMF for audio and speech applications, readers may refer to another chapter by Wang in this book.

SPARSE REPRESENTATION AND COMPRESSED SENSING

Separation of signals blindly from their underdetermined mixtures has attracted a great deal of attention over the past few years. It is a challenging source separation problem. One of the most common methods adopted for this problem is based on the sparse representation of signals (Zibulevsky & Bofill, 2001; Davies & Mitianoudis, 2004; Fevotte & Godsill, 2005; Zibulevsky & Pearlmutter, 2001). Closely related to sparse representation, there is an emerging technique called compressed sensing, which suggests that a signal can be perfectly recovered based on information rate, instead of the Nyquist rate, and random sampling, instead of uniform sampling, under certain conditions. It has been observed that compressed sensing exploits two important properties (Candès, 2006, Candès & Wakin, 2008; Donoho, 2006; Candès & Romberg, 2007). The first one is sparsity, which means that many natural signals can be represented in some proper basis in sparse (compressible) form. The second property is incoherence, i.e. the signal which is represented in some proper basis in sparse form should be dense as compared to the original representation of the signal. It is basically the extension of duality property between time and frequency domain.

There are similarities between the compressed sensing and source separation and their connections have been explored by (Blumensath & Davies, 2007), and further investigated by (Xu & Wang, 2008, Xu & Wang, 2009). It was found that the compressed sensing based signal recovery methods can be applied to the source reconstructions provided that the unmixing matrix is available or has been estimated (Zibulevsky & Bofill, 2001; Davies & Mitianoudis, 2004; Fevotte & Godsill, 2005; Zibulevsky & Pearlmutter, 2001; Blumensath & Davies, 2007).

A MULTISTAGE APPROACH

As mentioned above, both ICA and IBM have some limitations, i.e., the performance of the ICA is limited under the reverberant and noisy conditions and for the IBM technique, both the target speech and interference signal should be known *a priori*. In order to improve their performance, we have recently proposed a novel algorithm for the separation of convolutive speech mixtures based on the combination of ICA and IBM (Jan et al, 2009, Jan et al, 2010). The proposed method consists of

three steps. First, a constrained convolutive ICA algorithm (Wang et al., 2005) is applied to the binaural recordings to obtain the source signals. As is common to many other existing ICA algorithms, the separated target speech from this step still contains a considerable amount of interference from other sources. The performance steadily degrades with the increase of reverberation time (RT). To further reduce the interference, we use IBM to process the outputs from the previous step. Specifically, we estimate the ideal binary mask by comparing the energy of corresponding T-F units from the binaural outputs of the convolutive ICA algorithm. The estimated binary masks are then applied to the original mixtures for obtaining the target speech and interfering sources. The third step in our algorithm is to reduce musical noise using cepstral smoothing, where the noise was introduced by the errors in the estimation of the binary masks (Madhu et al., 2008; Araki et al., 2005). More specifically, we transform the binary mask into the cepstral domain, and smooth the transformed mask over time frames using the overlap-and-add technique. The benefit of doing this is that it is easier to distinguish the unwanted isolated random peaks from the mask patterns resulting from the spectral structure of the segregated speech in the cepstrum domain. As a result, we can apply different levels of smoothing to the binary T-F mask based on their various frequency ranges. The smoothed mask is transformed back into the T-F plane, which is then applied to the binaural outputs of the previous step in order to reduce the musical noise. Our multistage algorithm was first presented in (Jan et al., 2009), and the implementation details and systematic evaluations were provided in (Jan et al., 2010). Here, we briefly review this algorithm.

Stage 1. BSS of Convolutive Mixtures in the Frequency Domain

In a cocktail party environment, N speech signals are recorded by M microphones, and this can be described mathematically by a linear convolutive model,

$$x_i(n) = \sum_{i=1}^{n} \sum_{p=1}^{P} h_{ji}(p) s_i(n - p + 1) \quad (j = 1, \ldots, M) \tag{8}$$

where s_i and x_j are the source and mixture signals respectively, h_{ji} is a P-point room impulse response. This time-domain convolutive source separation problem can be converted to multiple instantaneous problems in the frequency domain (Wang et al., 2005; Araki et al., 2003) by applying short time Fourier transform (STFT). Using matrix notations, we have

$$X(k, m) = H(k) S(k, m) \tag{9}$$

where k represents the frequency index and m is the discrete time index. The mixing matrix H(k) is assumed to be invertible and time invariant. The sources are then estimated by apply an unmixing filter W(k) to the mixtures,

$$Y(k, m) = W(k) X(k, m) \tag{10}$$

where Y(k, m) represents the estimated source signals, and W(k) is estimated based on the assumption of independence. There are many algorithms that are suitable for this, e.g. (Araki et al., 2003; Parra & Spence, 2000; Sawada et al., 2007; Araki et al., 2007; Araki et al., 2004; Cichocki & Amari, 2002). In our multistage algorithm, we have used the constrained convolutive ICA approach in (Wang et al., 2005) for the separation in this stage. To further improve the separation quality, we apply the IBM technique to process the separated signal.

Stage 2. Combining Convolutive ICA and Binary Masking

Applying an inverse Fourier transform, Y(k, m) obtained above can be converted back to the time domain denoted as,

$$Y(n) = [Y_1(n)Y_2(n)]^T \tag{11}$$

Scaling is further applied to $Y_1(N)$and $Y_2(N)$ for obtaining the normalized outputs $\tilde{Y}_1(n)$ and $\tilde{Y}_2(n)$. After this we transform the two normalized outputs and into the T-F domain using STFT,

$$\tilde{Y}_i(k,m) = STFT(\tilde{Y}_i(n)) \tag{12}$$

By comparing the energy of each T-F unit of the above two spectrograms, the two binary masks are estimated as,

$$M_1^f(k,m) = \begin{cases} 1 \text{ if } \left|\tilde{Y}_1(k,m)\right| > \tau\left|\tilde{Y}_2(k,m)\right|, \\ 0 \text{ otherwise} \quad \forall \text{ k,m} \end{cases} \tag{13}$$

$$M_2^f(k,m) = \begin{cases} 1 \text{ if } \left|\tilde{Y}_2(k,m)\right| > \tau\left|\tilde{Y}_1(k,m)\right|, \\ 0 \text{ otherwise} \quad \forall \text{ k,m} \end{cases} \tag{14}$$

where τ is a threshold for controlling the sparseness of the mask, and typically $\tau = 1$ was used in our work (Jan et al, 2009, Jan et al, 2010). The masks are then applied to the T-F representation of the original two microphone recordings as follows

$$Y_i^f(k,m) = M_i^f(k,m).X_i(k,m) \quad i = 1,2 \tag{15}$$

The source signals in the time domain are recovered using the inverse STFT (ISTFT).

$$Y_i^t(n) = ISTFT(Y_i^f(k,m)) \quad i = 1,2 \tag{16}$$

As found in (Jan et al, 2009, Jan et al, 2010), this masking technique considerably improves the separation performance over that achieved by the convolutive ICA algorithm. However, a typical problem associated with the binary T-F masking is the so-called musical noise problem due to the the errors in mask estimation (Madhu et al., 2008; Araki et al., 2005). To address this issue, we employ a cepstral smoothing technique (Madhu et al., 2008) as detailed in the next subsection.

Stage 3. Cepstral Smoothing of the Binary Mask

The idea of using cepstral smoothing to reduce the musical artifacts was motivated by the speech production mechanism (Madhu et al., 2008; Oppenheim & Schafer, 1975). That is, for different frequency bands, different levels of smoothing are applied. By doing this, not only the broadband structure and pitch information in the speech signal are preserved, but also the musical artifacts can be reduced. Representing the binary masks of equation (6) and (7) in the cepstrum domain we have,

$$M_i^c(l,m) = DFT^{-1}\{ln(M_i^f(k,m)) \quad \{k = 0,...,K-1\} \tag{17}$$

where l and k are the quefrency bin index and the frequency bin index respectively (Madhu et al., 2008). *DFT* represents the discrete Fourier transform and K is the length of the DFT. After applying smoothing the resultant smoothed mask is given as,

$$\bar{M}_i^s(l,m) = \gamma_l \bar{M}_i^s(l,m-1) + (1-\gamma_l)M_i^c(l,m) \quad i = 1,2 \tag{18}$$

where γ_l is a parameter for controlling the smoothing level, and is selected as follows,

$$\gamma_l = \begin{cases} \gamma_{env} & if \quad l \in \{0,...,l_{env}\} \\ \gamma_{pitch} & if \quad l = l_{pitch} \\ \gamma_{peak} & if \quad l \in \{(l_{env}+1),...,K\} \backslash l_{pitch} \end{cases} \tag{19}$$

where $0 \le \gamma_{env} < \gamma_{pitch} < \gamma_{peak} \le 1$, l_{env} is the quefrency bin index that represents the spectral envelope of the mask and l_{pitch} is the quefrency bin index for the pitch. The underlying principle for the choice of γ_l is the illustrated as follows. $M^c(l,m)$, $l \in \{0,\ldots,l_{env}\}$, basically represents the spectral envelop of the mask $M^f(k,m)$. In this region the relatively low value is selected for γ_l to avoid the distortion in the envelope. Also, low smoothing is applied if l is equal to l_{pitch} so that the harmonic structure of the speech signal is maintained. High smoothing is applied in the last range to reduce the artifacts. Different from (Madhu ET Al., 2008) We calculate the pitch frequency by using the segregated speech signal obtained in the previous subsection, as follows,

$$l_{pitch} = argmax_l\{sig^c(l,m) \mid l_{low} \le l \le l_{high}\} \quad (20)$$

where $sig^c(l,m)$ is the cepstrum domain representation of the segregated speech signal $y^t(n)$. The range l_{low}, l_{high} is chosen so that it can accommodate pitch frequencies of human speech in the range of 50 to 500 HZ. The final smoothed version of the spectral mask is given as,

$$M_i^f(k,m) = \exp(DFT\{M_i^s(l,m) \mid l = 0,....,K-1\}) \quad (21)$$

This smoothed mask is then applied to the output segregated speech signals of the previous subsection to get the signals with reduced musical noise, as follows,

$$\bar{Y}_i^f(k,m) = M_i^f(k,m).Y_i^f(k,m) \quad (22)$$

RELATIONS TO OTHER METHODS

We have evaluated substantially the multistage approach discussed above using the audio mixtures generated by the simulated room model (Allen & Berkley, 1979), and the real recorded room impulse responses in (Pedersen et al., 2008). More details about the evaluations can be found in (Jan et al., 2009, Jan et al., 2010). Here, we only briefly discuss the separation performance that can be achieved with the multistage algorithm, as compared with two recent methods. Clean speech signals from a pool of 12 sources (Pedersen et al., 2008) were randomly selected to generate the reverberant mixture. In (Wang et al., 2005), the authors proposed a method for the segregation of speech signals using the frequency domain convolutive ICA approach. The results in terms of signal to noise ratio (SNR) for separated speech signals in (Wang et al., 2005) shows that the segregated signal contains a considerable amount of interference from other sources. In contrast to the method in (Wang et al., 2005), our proposed approach has better separation performance in terms of SNR measurements. Our results show that the multistage algorithm offers 3-4 dB gain in comparison to the method in (Wang et al., 2005). Listening tests also shows that our proposed method considerably improves the separation performance by suppressing the interference to a much lower level as compare to the method in (Wang et al., 2005). In the multistage algorithm, the complementary advantages of both techniques i.e. ICA and IBM are exploited to improve the performance of the separation system in contrast to the method in (Wang et al., 2005), and the musical noise is further reduced by cepstral smoothing.

The authors of (Pedersen et al., 2008) proposed a method in which ICA is also combined with IBM to improve the segregation performance in terms of interference suppression. However, our

multistage method employed cepstral smoothing which can reduce the artifacts (musical noise) introduced due to the estimation of the binary masks. Also a constrained convolutive ICA (Wang et al., 2005) is used in the multistage algorithm, while the instantaneous ICA algorithm is used in (Pedersen et al., 2008). It was also shown in (Jan et al., 2009, Jan et al., 2010) that the multistage approach is 18 times faster than the method in (Pedersen et al., 2008) in extracting the target speech signal from the convolutive mixture. Readers can find more details about the experimental set up and results including subjective listening test results of the multistage approach in (Jan et al., 2009, Jan et al., 2010).

APPLICATION AREAS

There are many applications that can benefit from the solution of the cocktail party problem, such as teleconferencing, speech recognition, bio-inspired systems, hearing aid, and reverberant environments, see e.g., (Yegnanarayana & Murthy, 2000, Wu & Wang, 2006). For example, in teleconferencing systems there might be multiple speakers talking at the same time, and echoes might also be a problem. To distinguish one speaker from another and the original speech from its echoes is necessary in this application. Progress in the cocktail party problem can facilitate the development of high quality teleconferencing systems with fewer practical constraints.

Speech recognition is another promising application area. Although the area of speech recognition has been developed for several decades and many successful systems have been implemented (John & Wendy, 2001; Junqua & Haton, 1995), the performance of these systems for uncontrolled natural environments is still limited. Any major progress in the cocktail party problem will prove to be crucial for the development of robust speech recognition systems that can deal with general auditory scenes within an uncontrolled natural environment.

As we have discussed, CASA is one of the most active areas of research for the cocktail party problem. In CASA, much effort has been devoted to the implementation (simulation) of the mechanism of the human auditory system. Similar ideas have evolved to the auditory scene analysis of non-human animals (Barker, 2006; Lippmann, 1997). Study of the cocktail party problem will facilitate our understanding of the designing techniques for the biologically inspired artificial scene analysis systems.

Research progress in cocktail party problem can be beneficial for other related applications in, for example, interference cancellation, deconvolution, and inverse problems. The common feature with these applications is that the propagation channels that the signals are transmitted are in multi path, and not known *a priori*, and is similar to what we have seen in a cocktail party environment. From this sense, the methods developed for the cocktail party problem are applicable for a broader area of applications.

CONCLUSION AND FUTURE RESEARCH

We have discussed the concept of the cocktail party problem, and in particular, the source separation issues in the cocktail party problem. We have presented several recent methods for speech source separation and auditory scene analysis, which are enabling techniques for addressing the cocktail party problem, including blind source separation, computational auditory scene analysis, non-negative matrix factorization, sparse representation, and model based techniques. Each method has its own advantages. As shown in the example, combinations of these techniques may achieve better separation performance. We have also briefly discussed the application areas of the

cocktail party processing. Future research may include reducing the room effect on an auditory mixture, dealing with the unknown number of sources and unknown type of sources, handling dynamic listening environment for multiple moving speakers, and analyzing the multimodal auditory data.

REFERENCES

Ainhoren, Y., Engelberg, S., & Friedman, S. (2008). *The cocktail party problem*. IEEE Instrumentation and Measurement Magazine.

Allen, J. & Berkley, D. (1979). Image method for efficiently simulating small-room acoustics, *J. of the Acoustical Soc. Am.* (pp. 943-950), 65(4).

Amari, S., Douglas, S. C., Cichocki, A., & Wang, H. H. (1997). Multichannel blind deconvolution and equalization using the natural gradient, *in Proc. IEEE Workshop Signal Process* (pp. 101–104).

Araki, S., Makino, S., Sawada, H., & Mukai, R. (2004). Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ICA. *In Proc. 5th International Conference Independent Component Anal. Blind Signal Separation* (pp. 898–905).

Araki, S., Makino, S., Sawada, H., & Mukai, R. (2005). Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask. *In Proc. IEEE International Conference Acoustics, Speech. Signal Processing*, *3*, 81–84.

Araki, S., Mukai, R., Makino, S., & Saruwatari, H. (2003). The fundamental limitation of frequency domain blind source separation for convolutive mixture of speech. *IEEE Transactions on Speech and Audio Processing*, *11*, 109–116. doi:10.1109/TSA.2003.809193

Araki, S., Sawada, H., Mukai, R., & Makino, S. (2007). Underdetermined blind sparse source separation for arbitrarily arranged multiple sources. In *EURASIP Journal App* (*Vol. 87*, pp. 1833–1847). Signal Process.

Arons, B. (1992). A review of the cocktail party effect. *Journal of the American Voice I/O Society,* 12, 35-50

Back, A. D., & Tosi, A. C. (1994). Blind deconvolution of signals using a complex recurrent network. In *Proc. IEEE Workshop Neural Networks Signal Process.* (pp. 565–574).

Barker, J. (2006). Robust automatic speech recognition. In Wang, D., & Brown, G. J. (Eds.), *Computational auditory scene analysis: Principles, algorithms, and applications* (pp. 297–350). Hoboken, NJ: Wiley-IEEE Press.

Blauert, J. (1983). *Spatial hearing: The psychophysics of human sound localization* (rev. Ed.). Cambridge, MA: MIT Press.

Blumensath, T., & Davies, M. (2007). Compressed sensing and source separation. In *International Conference on Independent Component Anal and Blind Source Separation.*

Bofill, P., & Zibulevsky, M. (2001). Underdetermined blind source separation using sparse representations. *Signal Processing*, 2353–2362. doi:10.1016/S0165-1684(01)00120-7

Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge, MA: MIT Press.

Bronkhorst, A. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple talker condition. *Acoustica*, *86*, 117–128.

Brown, G. J., & Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech & Language*, *8*(4), 297–336. doi:10.1006/csla.1994.1016

Buchner, H., Aichner, R., & Kellermann, W. (2004). Blind source separation for convolutive mixtures: A unified treatment. In Huang, Y., & Benesty, J. (Eds.), *Audio Signal Process. for Next-Generation Multimedia Communication Systems* (pp. 255–293). Boston, Dordrecht, London: Kluwer Academic Publishers. doi:10.1007/1-4020-7769-6_10

Candès, E. (2006). Compressive sampling. In *Proceedings of the International Congress of Mathematics*. Madrid, Spain. Candès, E. & Romberg, J. Sparsity and incoherence in compressive sampling. *Inverse Prob. 23*(3), 969–985

Candès, E. J., & Wakin, M. B. (2008). An introduction to compressive sampling. In *IEEE Signal Process Magazine.* (21).

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*, 975–979. doi:10.1121/1.1907229

Cichocki, A., & Amari, S. (2002). *Adaptive Blind Signal and Image Processing*. New York: Wiley Press. doi:10.1002/0470845899

Cichocki, A., Zdunek, R., & Amari, S. (2006). New algorithms for non-negative matrix factorization in applications to blind source separation. In *Proc. ICASSP* (Vol. 5, pp. 621–624) Toulouse, France.

Cooke, M., & Ellis, D. (2001). The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, *35*, 141–177. doi:10.1016/S0167-6393(00)00078-9

Cooke, M. P. (Dec 2002). Computational Auditory Scene Analysis in Listeners and Machines, *Tutorial at NIPS2002*, Vancouver, Canada.

Davies, M., & Mitianoudis, N. (2004, Aug). A simple mixture model for sparse overcomplete ICA. *IEE Proceedings. Vision Image and Signal Processing*, *151*(1), 35–43. doi:10.1049/ip-vis:20040304

Deng, L., & O'Shaughnessy, D. (2003). *Speech processing: A dynamic and optimization-oriented approach, ser. signal processing and communications*. London: Marcel Dekker, Taylor & Francis.

Donoho, D. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, *52*(4), 1289–1306. doi:10.1109/TIT.2006.871582

Douglas, S., Sawada, H., & Makino, S. (2005, Jan). Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters. *IEEE Transactions on Speech and Audio Processing*, *13*(1), 92–104. doi:10.1109/TSA.2004.838538

Douglas, S. C., & Sun, X. (2002). Convolutive blind separation of speech mixtures using the natural gradient. *Speech Communication*, *39*, 65–78. doi:10.1016/S0167-6393(02)00059-6

Feng, A. S., & Ratnam, R. (2000). Neural basis of hearing in real-world situations. *Annual Review of Psychology*, *51*, 699–725. doi:10.1146/annurev.psych.51.1.699

Fevotte, C., & Godsill, S. (2005). A bayesian approach for blind separation of sparse sources, In *IEEE Transactions on Speech and Audio Processing*. Washington D.C.

FitzGerald, D., Cranitch, M., & Coyle, E. (2005). Shifted non-negative matrix factorization for sound source separation. In *Proc. IEEE Int. Workshop on Statistical Signal Process.* Bordeaux, France (pp.1132-1137).

FitzGerald, D., Cranitch, M., & Coyle, E. (2006) Sound source separation using shifted non-negative tensor factorization. In *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process., 5*, 653-656.

Gygi, B., Kidd, G. R., & Watson, C. S. (2004). Spectral-temporal factors in the identification of environmental sounds. *The Journal of the Acoustical Society of America*, *115*(3), 1252–1265. doi:10.1121/1.1635840

Hall, D. E. (2001). *Musical Acoustics* (3rd ed.). Florence, Kentucky: Brooks Cole.

Haykin, S., & Chen, Z. (2005). The Cocktail Party Problem. *Journal Paper* [Cambridge, MA: MIT Press.]. *Neural Computation*, *17*, 1875–1902. doi:10.1162/0899766054322964

Hyvarinen, A., Karhunen, J., & Oja, E. (2001). *Independent Component Analysis*. John Wiley and Sons. doi:10.1002/0471221317

Ichir, M. H. & Djafari, A. M. (Jul 2006). Hidden markov models for wavelet based blind source separation, *IEEE Transaction on Image Process.* (pp. 1887–1899), vol. 15.

Jan, T. U., Wang, W., & Wang, D. L. (2009). A multistage approach for blind separation of convolutive speech mixtures. In *Proc ICASSP* (pp. 1713-1716). Taiwan.

Jan, T. U., Wang, W., & Wang, D. L. (2010). A multistage approach to blind separation of convolutive speech mixtures. In *IEEE Trans*. Audio Speech and Language Processing.

John, H., & Wendy, H. (2001). *Speech synthesis and recognition* (2nd ed.). London: Taylor & Francis.

Jourjine, A., Rickard, S., & Yılmaz, O¨. (2000). Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures. In *Proc. ICASSP* (Vol. 5, pp. 2985–8). Turkey.

Junqua, J. C., & Haton, J. P. (1995). *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. London: Kluwer Academic Publishers.

Kocinski, J. (2008). Speech intelligibility improvement using convolutive blind source separation assisted by denoising algorithms. In *EURASIP Journal*. *Speech Communication*, *50*, 29–37. doi:10.1016/j.specom.2007.06.003

Lambert, R. H., & Bell, A. J. (1997). Blind separation of multiple speakers in a multipath environment. In *Proc. IEEE International Conference Acoustics, Speech Signal Process.* (pp. 423–426).

Laurberg, H., Christensen, M. G., Plumbley, M. D., Hansen, L. K., & Jensen, S. H. (2008). *Theorems on positive data: on the uniqueness of NMF*. Computational Intelligence and Neuroscience.

Lee, D. D. & Seung, H. S. (1999). *Learning of the parts of object by non-negative matrix factorization, nature, 401*(10), 788-791.

Lee, D. D., & Seung, H. S. (2001). *Algorithms for non-negative matrix factorization. Advances in neural information processing* (pp. 556–562). Cambridge, MA: MIT Press.

Lee, T. W. (1998). *Independent Component Anal: Theory and Applications*. London: Kluwer Academic Publishers.

Lee, T. W., Bell, A. J., & Orglmeister, R. (1997). Blind source separation of real world signals. In *Proc. IEEE International Conference Neural Networks* (pp. 2129–2135).

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, *22*, 1–15. doi:10.1016/S0167-6393(97)00021-6

Madhu, N., Breithaupt, C., & Martin, R. (2008). Temporal smoothing of spectral masks in the cepstral domain for speech separation. In *Proc. ICASSP* (pp. 45–48).

Makino, S., Sawada, H., Mukai, R., & Araki, S. (2005). Blind source separation of convolutive mixtures of speech in frequency domain. In *IEICE Trans. Fundamentals*. *E (Norwalk, Conn.)*, *88-A*(7), 1640–1655.

Mango, K. N. (1991). *Hearing loss*. New York: Franklin Watts.

Matsuoka, K., & Nakashima, S. (2001). Minimal distortion principle for blind source separation. In *Proc. International Conference Independent Component Anal* (pp. 722–727), San Diego, CA, USA.

Mitianondis, N. & Davies, M. (2002). Audio source separation: solutions and problems. *International Journal of Adaptive Control and Signal Process.* (pp. 1–6).

Mukai, R., Sawada, H., Araki, S., & Makino, S. (2004) Frequency domain blind source separation for many speech signals. In *Proc. International Conference Independent Component Anal* (pp. 461-469).

Murata, N., Ikeda, S., & Ziehe, A. (2001, Oct). An approach to blind source separation based on temporal structure of speech signals. *Neuro Comput*, *41*(1-4), 1–24.

Naqvi, S. M., Zhang, Y., & Chambers, J. A. (2009). Multimodal blind source separation for moving sources. In *Proc ICASSP* (pp. 125-128), Taiwan.

Nickel, R. M., & Iyer, A. N. (2006). A novel approach to automated source separation in multispeaker environments. In *Proc. IEEE ICASSP* (pp. 629–632).

O'Shaughnessy, D. (2000). Speech communications-human and machin (2nd Ed.) In *Institute of electrical and electronic engineers*. New York.

Olsson, R. K., & Hansen, L. K. (2006). Blind separation of more sources than sensors in convolutive mixtures. In *Proc. IEEE ICASSP* (pp. 657–660).

Oppenheim, A. V., & Schafer, R. W. (1975). *Digital Signal Processing*. New Jersey: Prentice Hall.

Parra, L., & Spence, C. (2000). Convolutive blind separation of non stationary sources. *IEEE Transactions on Speech and Audio Processing*, *8*, 320–327. doi:10.1109/89.841214

Parry, R. M., & Essa, I. (2007). Incorporating phase information for source separation via spectrogram factorization. In *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process, 2,* 661-664. Honolulu, Hawaii.

Pearlmutter, B. A., & Zador, A. M. (2004). Monaural source separation using spectral cues. In *Proc. ICA 2004* (pp. 478–485).

Pedersen, M. S., Wang, D. L., Larsen, J., & Kjems, U. (2008). Two-microphone separation of speech mixtures. *IEEE Transactions on Neural Networks*, *19*, 475–492. doi:10.1109/TNN.2007.911740

Radfar, M. H., & Dansereau, R. M. (2007). Single channel speech separation using soft mask filtering. In *IEEE Trans. on Audio* (*Vol. 15*, pp. 2299–2310). Speech and Language Process.

Radfar, M. H., Dansereau, R. M., & Sayadiyan, A. (2006). Performance evaluation of three features for model-based single channel speech separation problem. In *Interspeech 2006, International Conference Spoken Language Process.* (ICSLP06), Pittsburgh, PA, (pp. 2610–2613).

Sawada, H., Araki, S., Mukai, R., & Makino, S. (2007). Grouping separated frequency components by estimating propagation model parameters in frequency domain blind source separation. In *IEEE Transaction Speech Audio Language Process.* (Vol. 15, pp. 1592–1604).

Sawada, H., Mukai, R., Araki, S., & Makino, S. (2003). Polar coordinate based nonlinear function for frequency-domain blind source separation. In *IEICE Transactions Fundamentals*, E86 (3), 590–596.

Sawada, H., Mukai, R., Araki, S., & Makino, S. (2004). A robust and precise method for solving the permutation problem of frequency domain blind source separation. *IEEE Transactions on Speech and Audio Processing*, *12*, 530–538. doi:10.1109/TSA.2004.832994

Schmidt, M. N., & Laurberg, H. (2008). *Non-negative matrix factorization with Gaussian process priors*. Computational Intelligence and Neuroscience.

Schmidt, M. N., & Morup, M. (2006). Nonnegative matrix factor 2D deconvolution for blind single channel source separation. In *Proc. 6th Int. Conf. on Independent Component Analysis and Blind Signal Separation*. Charleston, SC, USA, (pp. 700-707).

Schmidt, M. N., & Olsson, R. K. (2006). *Single-channel speech separation using sparse non-negative matrix factorization*. Interspeech.

Schobben, L., & Sommen, W. (2002). A frequency domain blind signal separation method based on decorrelation. *IEEE Transactions on Signal Processing*, *50*(8), 1855–1865. doi:10.1109/TSP.2002.800417

Smaragdis, P. (2004). Non-negative matrix factor deconvolution, extraction of multiple sound sources from monophonic inputs. In *Proc. 5th Int. Conf. on Independent Component Analysis and Blind Signal Separation*. Granada, Spain, (LNCS 3195, pp.494-499).

Smaragdis, P. (2007). Convolutive speech bases and their application to supervised speech separation. In *IEEE Trans. Audio Speech and Language Processing, 15*(1), 1-12.

Smaragdis, P., & Brown, J. C. (2003). Nonnegative matrix factorization for polyphonic music transcription. In *IEEE Int. Workshop on Applications of Signal Process. to Audio and Acoustics*, New Paltz, NY. (pp. 177-180).

Soon, V. C., Tong, L., Huang, Y. F., & Liu, R. (1993). A robust method for wideband signal separation. In *Proc. IEEE International Symposium Circuits Systems* (Vol.1, pp. 703–706).

Todros, K., & Tabrikian, J. (2004) Blind separation of non stationary and non gaussian independent sources. In *Proc. IEEE Convention of Electrical and Electronics in Israel.*

University of Sheffield, Department of Computer Science. (2007). *Auditory scene analysis: Listening to several things at once.* Retrived on June 27,2009 from http://www.dcs.shef.ac.uk/spandh/research/asa.html

Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. In *IEEE Trans. Audio, Speech, and Language Process, 15,* (3).

Wang, B., & Plumbley, M. D. (2005). Musical audio stream separation by non-negative matrix factorization. In *Proc. DMRN Summer Conference*. Glasgow, UK.

Wang, D. L. (2005). On ideal binary mask as the computational goal of auditory scene analysis. In Divenyi, P. (Ed.), *Speech Separation by Humans and Machines* (pp. 181–197). Norwell, MA: Kluwer Academic. doi:10.1007/0-387-22794-6_12

Wang, D. L., & Brown, G. J. (1999, May). Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks*, *10*, 684–697. doi:10.1109/72.761727

Wang, D. L., & Brown, G. J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley/IEEE Press.

Wang, W. (2007). Squared Euclidean distance based convolutive non-negative matrix factorization with multiplicative learning rules for audio pattern separation. In *Proc. IEEE Int. Symp. on Signal Proces. and Info. Tech.* Cairo, Egypt.

Wang, W., Cichocki, A., & Chambers, J. A. (2009). A multiplicative algorithm for convolutive non-negative matrix factorization based on squared Euclidean distance. In *IEEE Trans* (pp. 447–452). On Signal Processing.

Wang, W., Luo, Y., Sanei, S., & Chambers, J. A. (2008). Note onset detection via non-negative factorization of magnitude spectrum, In *EURASIP Journal on Advances in Signal Processing* (pp. 447-452).

Wang, W., Sanei, S., & Chambers, J. A. (2005). Penalty function-based joint diagnolization approach for convolutive blind separation of nonstationary sources. *IEEE Transactions on Signal Processing*, *53*, 1654–1669. doi:10.1109/TSP.2005.845433

Wu, M., & Wang, D. L. (2006). A two-stage algorithm for one-microphone reverberant speech enhancement. In *IEEE Transaction on Audio, Speech, and Language Process, 14.*

Xu, T., & Wang, W. (2009). A compressed sensing approach for underdetermined blind audio source separation with sparse representations. In *Proc. IEEE Int. Workshop on Statistical Signal Processing*. (pp. 493-496). Cardiff, UK.

Xu, T., & Wang, W. (2010). A block-based compressed sensing method for underdetermined blind speech separation incorporating binary mask. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. Texas, USA.

Yegnanarayana, B., & Murthy, P. S. (2000, May). Enhancement of reverberant speech using LP residual signal. *IEEE Transactions on Speech and Audio Processing*, *8*(3), 267–281. doi:10.1109/89.841209

Yost, W. A. (1997). The cocktail party problem: Forty years later. In Gilkey, R., & Anderson, T. (Eds.), *Binaural and spatial hearing in real and virtual environments* (pp. 329–348). Ahwah, NJ: Erlbaum.

Yost, W. A. (2000). *Fundamentals of hearing: An introduction* (4th ed.). San Diego: Academic Press.

Zibulevsky, M., & Bofill, P. (2001). Underdetermined blind source separation using sparse representations. *Signal Processing*, *81*(11), 2353–2362. doi:10.1016/S0165-1684(01)00120-7

Zibulevsky, M., & Pearlmutter, B. A. (2001). Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, *13*(4), 863–882. doi:10.1162/089976601300014385