This item was submitted to [Loughborough's Research Repository](#) by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

Adaptive blind moving source separation based on intensity vector statistics

PLEASE CITE THE PUBLISHED VERSION

PUBLISHER

Elsevier B.V.

VERSION

AM (Accepted Manuscript)

PUBLISHER STATEMENT

This paper was accepted for publication in the journal Speech Communication and the definitive published version is available at <https://doi.org/10.1016/j.specom.2019.08.001>.

LICENCE

CC BY-NC-ND 4.0

REPOSITORY RECORD

Riaz, Areeb, Xiyu Shi, and Ahmet Kondo. 2019. "Adaptive Blind Moving Source Separation Based on Intensity Vector Statistics". figshare. <https://hdl.handle.net/2134/9536765.v1>.

Adaptive Blind Moving Source Separation Based on Intensity Vector Statistics

Areeb Riaz^{a,1}, Xiyu Shi^{b,*}, Ahmet Kondo^z^b

^a*Department of Electrical and Electronic Engineering, University of Surrey, Guildford, GU2 7XH, UK*

^b*Institute for Digital Technologies, Loughborough University London, E15 2GZ, UK*

Abstract

This paper presents a novel approach to blind moving source separation by detecting, tracking and separating speakers in real-time using intensity vector direction (IVD) statistics. It updates unmixing system parameters swiftly in order to deal with the time-variant mixing parameters. Denoising is carried out to extract reliable speaker estimates using von-Mises modeling of the IVD measurements in space and IIR filtering of the IVD distribution in time. Peaks in the IVD distribution are assigned location expectation values to check for consistency, and consequently high location expectation peaks are declared as active speakers. The location expectation algorithm caters for natural pauses during speech delivery. Speaker movements are tracked by spatial isolation of the detected peaks using time-variant regions of interest. As a result, the proposed moving source separation system is capable of blindly detecting, tracking and separating moving speakers. A real-time demonstration has been developed with the proposed system pipeline, allowing users to listen to active speakers in any desired combination. The system has an advantage of using a small coincident microphone array to separate any number of moving sources utilising the first order Ambisonics signals while assuming source signals to be W-disjoint orthogonal. Being nearly closed-form, the proposed system does not require convergence or initialization of parameters.

*Corresponding author

Email addresses: areeb@goertek.eu (Areeb Riaz), x.shi@lboro.ac.uk (Xiyu Shi), a.kondo@lboro.ac.uk (Ahmet Kondo)

¹Present address: Sundkaj 9, 1 TH, Nordhavn, 2150 Copenhagen, Denmark.

Keywords: Blind moving source separation, convolutive mixtures, intensity vector direction, location expectation, speaker activity detection

1. Introduction

Acoustic blind moving source separation (BMSS) is the problem of automated detection, tracking and separation of moving sources from the observed mixtures. While humans have the ability to focus on a speaker of interest among several active speakers, for machines the problem is still being widely investigated. The blind source separation (BSS) problem was first defined by Cherry (1953). It is termed “blind” due to the absence of any prior information which may assist in performing the separation.

Considerable research has been carried out in the context of physically stationary sources (Khan et al., 2013; Sturmel and Daudet, 2013; Souden et al., 2013; Sawada et al., 2011; Bao et al., 2013; Yoshioka et al., 2011; Mandel et al., 2010), while moving sources have been notably less investigated (Traa and Smaragdis, 2014; Naqvi et al., 2010; Koldovsky et al., 2013; Chong et al., 2014a). Movement of speakers makes the source separation problem more challenging due to time-variant mixing parameters. Unmixing parameters cannot be accurately and swiftly measured due to lack of sufficient data over which the mixing conditions remain the same (Naqvi et al., 2012). In real-world scenarios, the BMSS problem is multifaceted. It includes swift and accurate approximation of the room impulse response (RIR), speaker activity detection (in case a speaker enters or exits the acoustic scene), reliable tracking of speaker movements, practical geometry of the microphone array and computational complexity of the algorithm.

The BMSS problem for M microphones and N sources can be mathematically represented as:

$$X(t) = H(t)S(t), \quad (1)$$

where the N source signals, $S(t)$, are represented as $[s_1(t)s_2(t)s_3(t)...s_N(t)]^T$, assuming $s_n(t)$ is the n th source at time t . The transfer function coefficients are contained in the $(M \times N)$ time-variant mixing matrix, $H(t)$, such that $h_{mn}(t)$ represents the coefficient for m th microphone and n th source. $X(t) = [x_1(t)x_2(t)x_3(t)...x_M(t)]^T$ contains the resultant mixture components of M microphone, where $x_m(t)$ represents the mixture signal recorded by m th microphone at time t . The task of any BMSS system is to extract

source signals $S(t)$ from the observed mixtures $X(t)$. For physically stationary sources the dependency of transfer function, H , on time, t , is absent unless the surrounding environment changes.

In practical scenarios, the separation of moving acoustic sources is a very useful pre-processing stage for speaker identification and speech recognition problems (Yamamoto et al., 2006; Koutvas et al., 2000). For example, a speaker may move while delivering a presentation. Similarly, other applications related directly to BMSS are speaker activity detection, speech enhancement, automatic camera control, high quality hearing aids and teleconferencing (Plinge et al., 2012; Pulkki, 2007).

The existing literature on this subject proposes several BSS algorithms. Among deterministic approaches, a nearly closed-form IVD-based algorithm proposed by Gunel et al. (2008) exploits space-time-frequency sparsity among source signals to perform separation. The IVD-based algorithm utilises a compact, nearly coincident microphone array to carry out intensity measurements. The nearly closed-form nature of the IVD measurement leads to a computationally efficient system. It can perform low delay real-time separation, and the number of sources to be separated is not restricted to the number of obtained mixtures. However, the IVD-based system in (Gunel et al., 2008) is incapable of dealing with dynamically changing mixing conditions with moving sources.

In this paper, we propose a novel BMSS system, based on the IVD statistics, capable of detecting, tracking and separating multiple moving speakers. In Section II, related work in the field of source separation is discussed. Section III presents the derivation of IVD measurement and the IVD distribution representation in space. Section IV analyses the IVD distribution in space to examine and understand the behaviour and potential of the IVD-based system with respect to different practical aspects. Section V explains the methodology of the proposed system with implementation details. Section VI evaluates the proposed BMSS system and discusses experimental results. Section VII concludes the publication.

2. Related Work

The majority of existing source separation algorithms assume physically stationary sources. In practical acoustic scenarios, one or more speakers may move in the surrounding space and/or the number of speakers may

vary with time. As a result, data inadequacy for accurately estimating time-variant unmixing parameters limits the performance of most existing source separation algorithms.

Stochastic algorithms, such as Independent Component Analysis (ICA), perform source separation by assuming statistical independence among source signals (Comon and Jutten, 2010; Cardoso, 1998). ICA-based algorithms approximate the inverse of the mixing matrix to extract source signals from the obtained mixtures. The algorithms require several iterations for deriving the unmixing matrix, leading to higher computational costs. Much faster implementations also exist, such as the FastICA; however, convergence is still required (Hyvarinen, 1999; Douglas et al., 2007). Even after successful separation of the source signals using ICA, permutation and scaling problems arise while reconstructing the source signals. As each frequency bin is considered an instantaneous mixture, the permutation problem is more destructive in the frequency domain than in the time domain.

The permutation problem adds to the computational complexity as post-processing steps need to be carried out (Smaragdis, 1998; Mitianoudis and Davies, 2004), and as no closed-form solution exists, this is still being widely investigated (Sarmiento et al., 2015; Zhao et al., 2014). In similar work with moving speakers, visual cues are used to help solve the permutation problem using 3D video trackers. As a result, visual source tracking, beam-forming and intelligently initialized FastICA are used together to perform source tracking and separation (Naqvi et al., 2010). In other ICA-based work, semi-blind source separation is performed using prior knowledge of the cancellation filters (CF) for a target region in space (Koldovsky et al., 2013). Fundamentally, stochastic algorithms alone cannot adequately estimate time-variant unmixing filter coefficients for moving speaker convolutive mixtures in a blind manner. As such, only a handful of publications can be found in the field of BMSS, either assuming instantaneous mixtures or the mixing coefficients varying sufficiently slowly (Naqvi et al., 2010; Anemüller and Gramssý, 1999; Addison and Roberts, 2006; Prieto and Jinachitra, 2005), and almost all the work has been done considering two speakers.

Adaptive algorithms, such as the Adaptive beam-forming (ABF), extract a desired source and suppress the undesired interferences from convolutive mixtures by applying spatial selectivity to multichannel filters (Griffiths and Jim, 1982; Parra and Alvino, 2002). In ABF, the beam-forming filter’s coefficients are initialised with the microphone array geometry information and optimized based on the signal properties. Like stochastic algorithms, adap-

tive algorithms need to converge to a solution and the convergence depends on source signal properties. When the source signals are ideally independent, two sets of frequency domain ABF have been proven to be equivalent to the frequency domain BSS (Araki et al., 2003a). The ABF needs only the jammer signal to be active for adaptation, whereas frequency domain BSS can adapt when both the target and jammer signals are active. Also, ABF loses effectiveness with null beam-forming for interfering sources in more reverberant environments (Araki et al., 2003b).

Sparsity-based algorithms, such as the Degenerate Unmixing Estimation Technique (DUET), depend on completely different aspects. The Spatial signature of sources and the multipath characteristics of the reverberant environment are exploited to perform deterministic sparsity-based source separation (Van Der Veen, 1998; Shindo and Hirai, 2002; Yamashita et al., 2004). DUET exploits relative amplitude and delay differences between the microphones to cluster similar time-frequency (T-F) components (Rickard, 2007). More recent work utilising DUET employs EM initialised by random sample consensus (RANSAC) to cluster wrapped inter-channel phase differences (Traa and Smaragdis, 2014). For moving speakers, with RANSAC pre-processed data, the factorial wrapped Kalman filter is used to perform speaker tracking; however, the number of speakers is assumed to be known *a priori* (Traa and Smaragdis, 2014). Mandel et al. (2010) proposed a model-based expectation maximization (EM) source separation and localization (MESSL) algorithm. Each source in a mixture is described by a probabilistic model of interaural parameters, namely the interaural level difference (ILD) and interaural phase difference (IPD). Individual spectrogram points based on their ILDs and IPDs are grouped to form masks to isolate individual sound sources. Similar method is also used in (Alinaghi et al., 2014). However, the performance of MESSL algorithm suffers when the sources are close to each other with small angular displacement.

Günel et al. (2008) first proposed a sparsity-based source separation algorithm using a compact, nearly coincident microphone array. Prior to this, the coincident microphone arrays were mainly used for intensity measurement (Bree et al., 1999), sound source localization (Günel et al., 2007) and directional audio coding (Merimaa and Pulkki, 2005). As such, non-coincident microphone arrays are used in almost all the source separation techniques, including ICA, ABF, MESSL and DUET (Traa and Smaragdis, 2014; Araki et al., 2003b; Mandel et al., 2010; Naqvi et al., 2012). The use of non-coincident microphone array leads to an inaccurate approximation of the

multipath characteristics as the channel characteristics are different at each microphone.

Similar work has been done by Chong et al. (2014a,b) in source detection and tracking, by using a combination of DUET and a cardinality balanced multi-target multi-Bernoulli filter; however, this work has only considered simulated mixture data with only up to two sources, without performing the source separation stage.

Huang et al. (2014) first introduced deep neural networks (DNNs) and recurrent neural networks (RNNs) to speech source separation in a supervised setting (Wang and Chen, 2018). The networks do not use the spectral representation directly for source separation; rather they learn the optimal hidden representations of mixtures through multiple layers of non-linearity and reconstruct the spectral domain signals based on the learned hidden representations. Yu et al. (2016) has used DNNs for stereo speech source separation, where the likelihood of a T-F unit belonging to a particular source is estimated and a soft T-F mask is generated for separation. Combined with an iterative EM algorithm, multiple DNNs are also used to improve the spectral estimation to derive a multichannel filter for source separation (Nugraha et al., 2016a,b). Uhlich et al. (2017) has used data augmentation to avoid over-fitting in the training of DNNs and RNNs, which further improved the performance of neural networks for music separation. All of these networks need considerable training to learn the spectral features of the sources. Note that the neural networks technique has not yet been used for separating moving sources.

To the best of our knowledge, no convolutive source separation system has previously demonstrated the ability to detect, track and separate up to four moving speakers.

3. Theoretical Background

The intensity of a sound field in a specified direction at a point is the sound energy transmitted per unit of time in the specified direction through a unit area normal to this direction at the point. The intensity, I , of a plane wave is

$$I = P^2/(\rho c), \quad (2)$$

where P is the pressure, ρ is density of the medium and c is the velocity of propagation (Olson, 1957, pp.15-16). This implies that if the pressure

is obtained as a vector quantity, the intensity as well as the direction of an acoustic plane wave can be measured. The first order Ambisonics, also known as the B-format signals, can be used to get the pressure and its projection along the Cartesian axes to measure the intensity as a vector.

For a sound field along the horizontal plane, B-format signals comprise of $w(t)$, $x(t)$, and $y(t)$ signals, where $w(t)$ is the omnidirectional pressure, and, $x(t)$ and $y(t)$ are the pressure components along the x -axis and y -axis respectively. The STFT spectra of B-format signals can be represented as (Benjamin and Chen, 2005):

$$W(\omega, t) = 1/\sqrt{2} \times P(\omega, t), \quad (3)$$

$$X(\omega, t) = \cos(\theta) \times P(\omega, t), \quad (4)$$

$$Y(\omega, t) = \sin(\theta) \times P(\omega, t), \quad (5)$$

where θ denotes the horizontal angle of the incoming acoustic wave at a particular T-F unit with respect to the orientation of the microphone array, and $P(\omega, t)$ denotes the STFT spectra of sound pressure.

Thus, using the STFT spectra of B-format signals the intensity vector along the horizontal plane can be represented as:

$$\vec{I}(\omega, t) = \frac{1}{\rho c} (\mathbf{Re}\{W^*(\omega, t) \cdot X(\omega, t)\} + i\mathbf{Re}\{W^*(\omega, t) \cdot Y(\omega, t)\}), \quad (6)$$

where $*$ denotes complex conjugate, $\mathbf{Re}\{\cdot\}$ denotes taking the real part of its argument, and i means the imaginary unit (Gunel et al., 2008). The IVD measurement at a particular T-F unit can thus be determined using (6) as:

$$IVD(\omega, t) = \tan^{-1} \left(\frac{\mathbf{Im}\{\vec{I}(\omega, t)\}}{\mathbf{Re}\{\vec{I}(\omega, t)\}} \right), \quad (7)$$

where $\mathbf{Im}\{\cdot\}$ denotes taking the imaginary part of its argument.

In the literature, the Laplacian and von-Mises distributions have been used to model sparse data in space (Gunel et al., 2008; Traa and Smaragdis, 2013; Mitianoudis, 2012). When considering reverberation, we have modelled each IVD measurement with a von-Mises distribution function (Mardia and Jupp, 2009, pp.36–45) as follows:

$$f(\theta; IVD(\omega, t), \sigma) = \frac{e^{\sigma \cos(\theta - IVD(\omega, t))}}{2\pi I_o(\sigma)}, \quad (8)$$

where σ is the concentration parameter, similar to variance of a Gaussian distribution, θ is a circular variable lying within 0 and 2π , and $I_o(\sigma)$ is the modified Bessel function of order zero. σ is logarithmically related to the 6 dB beam-width θ_{BW} as follows:

$$\sigma = \ln 2 / [1 - \cos(\theta_{BW}/2)]. \quad (9)$$

θ_{BW} should be set as a small value to maintain localization accuracy. For each STFT frame, using von-Mises modelling, the IVD distribution in the surrounding space is measured as:

$$H(IVD; (\omega, t)) = \sum_{IVD=1}^{360} f(\theta; IVD(\omega, t), \sigma). \quad (10)$$

4. Analysis of IVD Distribution in Space

In order to understand how an IVD-based source separation system deals with realistic acoustic scenarios, we will investigate the distribution of IVD in space in this section, and will present and discuss the relevant observations. All observations have been made with convolutive recordings of human subjects in an ITU-R BS1116 standard listening room with a reverberation time of 320ms. This analysis has been carried out using STFT with the overlap-and-add (OLA) technique of 2048 samples window length and 50% overlap at 44100 Hz sampling frequency.

4.1. Acoustic Source Activity

Fig.1 shows the IVD distribution of two physically stationary speakers over a period of 2.5s. The x , y and z axes represent the time frames, the surrounding space and the IVD distribution, respectively. The speakers are active at 45 and 135 degrees with respect to the microphone array. The unprocessed IVD distribution in Fig.1a can be seen to have noisy information of the two speakers. Reflections and reverberations, along with W-disjoint orthogonality violation contribute to such an observation. Also speech signals by nature are sparse in the frequency domain and impulsive at times, and as a result, the T-F units corresponding to the same speaker may be located away from each other by a small angular interval. Unused T-F components behave

Figure 1: IVD Distribution in space (a) Unprocessed IVD distribution. (b) Applying first order IIR filter to denoise in time domain. (c) Modelling the IVD with von-Mises distribution function to denoise in space. (d) Applying von-Mises model along with the first order IIR filter to achieve denoised IVD distribution in both time and space.

like random noise under the effect of spectral leakage due to processing with windowed DFT. The frequency dependent environment transfer function also dictates the varying IVD measurement across different frequencies. In order to obtain reliable speaker information, a first order infinite impulse response (IIR) filter is applied to denoise the IVD distribution.

The effect of IIR filtering depicted in Fig.1b represents 200ms of time averaging. As a result, a more distinguished and prominent IVD distribution of each speaker is achieved. Similar to the first order IIR filtering, which caters for noise in the time domain, denoising in space is achieved by modelling each IVD as a von-Mises distribution function (Fig.1c). Denoising in both time and space results in a robust IVD distribution representation of the speakers

Figure 2: Effect of speaker distance on IVD distribution in space. Speaker 1 is moving backward and forward w.r.t to microphone array while Speaker 2 is stationary in space.

(Fig.1d), and can be used to reliably detect and track speaker movements over short time durations.

4.2. Speaker Distance to Microphone Array

The distance of a speaker to the microphone array placed at the center of the room dictates two parameters: speaker energy and its direct to reverberant ratio (DRR). Fig.2 depicts the scenario of a speaker which varies its distance to the microphone array while another physically stationary speaker is active in the vicinity. Speaker 2 is at a constant distance of 1m to the microphone array while Speaker 1 starts at 1m, comes as close as 0.3m, and moves away to 1.6m, facing towards the microphone array all the time.

As Speaker 1 comes close to the microphone array, due to an increase in overall speaker energy, comparatively low energy T-F components also start to contribute to the speaker IVD distribution. This is reflected by a rise in the IVD distribution of Speaker 1 (highlighted within red lines in Fig.2). It should also be noted that at time frames when Speaker 1 is closest to the microphone array (Fig.2, time frames 200 to 300), Speaker 2 is observed to have the lowest IVD distribution. As Speaker 1 moves away from the reflective walls and closer to the microphone array, the DRR gets

Figure 3: IVD distribution of two moving speakers at the same distance to the microphone array. Speaker 1 is moving between 75 and 125 degrees, while Speaker 2 is moving between 175 and 225 degrees.

higher as well. Consequently, a more prominent IVD distribution in space is also achieved because of the reduced impact of reflections on the direct sound components. As Speaker 1 starts moving away from the microphone array, the IVD distribution gets lower (Fig.2, time frame 300 onwards) and eventually Speaker 2 is more prominent when the distance of Speaker 1 is greater than that of Speaker 2. As such, a distant low energy speaker may not be detected using the IVD statistics and unfavourably acts as diffused background babble noise.

Given that both Speaker 1 and Speaker 2 have an approximately similar active speech level, it can be said that the IVD distribution is dependent on the relative energy of the speakers as well as their absolute energy. Also the DRR should be sufficiently high to exploit the sparsity-based IVD statistics for performing BMSS in an effective manner.

4.3. Moving Acoustic Sources

With speakers moving in the surrounding environment, the IVD distribution is analysed over short time durations in order to perform speaker tracking. The denoised IVD distribution of two moving speakers over a pe-

Figure 4: Block diagram of the proposed BMSS System

Figure 5: Tetrahedral configuration of cardioid microphones

riod of 15s is depicted in Fig.3. The speaker movements can be seen to take place in two distinct paths, highlighted within white boundary lines. Over this period of time, the speakers are either facing sideways or directly towards the microphone array, leading to limited variations in the captured energy and DRR. The change in location of a speaker over time can be seen as a time-varying IVD distribution peak in its respective region.

Considering such denoised observations of time-variant IVD distributions in practical acoustic scenarios, we propose a novel BMSS system in the next section to perform speaker detection, tracking and separation by reliably and swiftly estimating, and updating the unmixing parameters.

5. Proposed Methodology

The block diagram of the proposed BMSS system is shown in Fig.4. This section describes various algorithms in the blocks.

5.1. A-Format Audio

The sound field has been captured by an A-format microphone array, namely the Core Sound TetraMic microphone (Core-sound.com, 2017). In

comparison with microphone arrays for beamforming, the A-format microphone has a small footprint, the performance for low frequencies does not depend on greater spacing between the microphones, and it is commercially available. The A-format signals can be conveniently converted into the required B-format signals with (3-5), i.e. W the omnidirectional pressure of the sound field, and X , Y the pressure gradients along the x -axis and y -axis, for spatial intensity measurements. This makes sources tracking and localization relatively easier by analyzing the sound direction of arrival from intensity vectors.

The TetraMic consists of four cardioid microphones facing towards the non-adjacent ends of a cube. The placement of these microphones is shown in Fig.5, where LFU , RFD , LBD and RBU are the left-forward-up, right-forward-down, left-back-down and right-back-up cardioid microphones with respect to the center of the cube. Such tetrahedral configuration of cardioid microphones covers the surrounding space in 3D. Collectively the four cardioid microphone signals are known as A-format.

5.2. Frequency Domain Representation

It is assumed that the source signals are W-disjoint orthogonal, mathematically represented as:

$$s_i(\omega, t) \times s_j(\omega, t) = 0, \quad (11)$$

where $s_i(\omega, t)$ and $s_j(\omega, t)$ are the T-F content from i th and j th sources respectively. In reality speech signals are nearly W-disjoint orthogonal (Rickard and Yilmaz, 2002), and this assumption tends to become more violated with a greater number of sources and increasing reverberation time. However, to some extent, the sources can still be separated, but with reduced unmixing performance (Traa and Smaragdis, 2013; Yilmaz and Rickard, 2004). Holding onto this basic assumption, transformation is made into the frequency domain using the STFT technique. A-format data are windowed via a sine function, given as:

$$w_k = \sin \left(\frac{\pi}{2} \sin^2 \left[\frac{\pi}{2M} \left(k + \frac{1}{2} \right) \right] \right), \quad (12)$$

where $2M$ is the window length in samples and k is the sample index. An overlap of 50% is maintained between the subsequent time windows to carry out OLA (Princen and Bradley, 1986). As compared to the commonly used Hamming or Hanning window (Heinzel et al., 2002), the given sine window has a lower side lobe level.

5.3. A-Format to B-Format Conversion

To obtain the pressure and pressure gradient signals, A-format signals are converted into B-format as follows (Benjamin and Chen, 2005):

$$W(\omega, t) = P_{LFU}(\omega, t) + P_{RFD}(\omega, t) + P_{LBD}(\omega, t) + P_{RBU}(\omega, t), \quad (13)$$

$$X(\omega, t) = P_{LFU}(\omega, t) + P_{RFD}(\omega, t) - P_{LBD}(\omega, t) - P_{RBU}(\omega, t), \quad (14)$$

$$Y(\omega, t) = P_{LFU}(\omega, t) - P_{RFD}(\omega, t) + P_{LBD}(\omega, t) - P_{RBU}(\omega, t), \quad (15)$$

where P_{LFU} , P_{RFD} , P_{LBD} and P_{RBU} are the observed pressures of the LFU, RFD, LBD and RBU cardioid microphones, respectively.

5.4. IVD Measurement

Using B-format signals, the IVD is measured at each T-F unit using (7). This measurement is dependent on several factors including the microphone array response, reverberant response of the surrounding environment, relative energy of the speakers, their DRR, the impact of multiple speakers on a particular T-F unit and the physical aperture of the sound sources. As a result the IVD measurements are spatially spread around the physical locations of the speakers.

5.5. IVD Distribution

Several acoustic scenarios have been discussed in Section 4, which depict system behaviour considering IVD distribution in space. In a situation with simultaneously active speakers, the greater the number of time frames used for IVD distribution representation, the greater the confidence in speaker locations; however, this is no longer true when the spatial signatures of speakers are time-varying. For moving speakers, it is critical to have the ability to detect speaker locations as well as track their movements over short time durations.

It has also been observed that the system's ability to perform speaker detection is dependent on the considered frequency range. For speech signals, most of the energy lies in vowels and consonants up to 4000 Hz and unvoiced consonants are mostly present up to 8000 Hz. Hence, an analysis frequency range which can reflect this provides better detection results. The IVD distribution in a two-speaker scenario with an increase in the considered frequency range is depicted in Fig.6. The T-F components up to 4000 Hz

Figure 6: IVD distribution with different source frequency ranges. (a) 80 - 4000 Hz. (b) 80 - 16000 Hz.

(Fig.6a) and 16000 Hz (Fig.6b) have been shown in comparison. Speaker 2, distant to the microphone array (and comparatively low in energy), portrays less prominent IVD distribution, and this becomes even less at a higher frequency range. As the nature of speech is sparse in the T-F domain, excessively unused T-F components lead to such an observation. Hence an analysis frequency range which is relevant to the signals is critical in giving reliable source estimates.

While processing one time frame to another, IIR filtering of the IVD distribution is carried out to denoise in the time domain as:

$$H(IVD; (\omega, t)) = \alpha \cdot H(IVD; (\omega, t)) + (1 - \alpha) \cdot H(IVD; (\omega, t - 1)), \quad (16)$$

where the IVD distribution of the current frame, $H(IVD; (\omega, t))$, is updated using the filtered IVD distribution of the previous frame, $H(IVD; (\omega, t - 1))$, based on the IIR filter coefficient α . The parameter α was empirically determined to be corresponding to an averaging time between 180ms and 250ms for sufficient accuracy and reliable speaker estimates in our experiments, where both the location of the speakers and their numbers are time-variant. This was done by using labeled recordings for true speaker activity and then matching them with speaker detection results obtained from the designed algorithm. The IVD distributions and results of peak detections of two moving speakers with α equal to 0.5, 0.125 and 0.0625 characterizing 46.4ms, 185.8ms and 371.5ms time averaging are shown in Fig.7.

Figure 7: (a), (c) and (e) IIR filtered IVD distributions with coefficient $\alpha = 0.5, 0.125$ and 0.0625 respectively. (b), (d) and (f) Results of peak detection and location expectation algorithm with IVDs in (a), (c) and (e) respectively.

Figure 8: Average IVD distribution of stationary speakers over 2.5 seconds. (a) Average IVD distribution peaks are higher and more prominent with two speakers. (b) IVD noise floor rises due to greater reverberations and overlapping time-frequency data with a greater number of speakers.

Fig.7a shows that the IVD distribution peaks are less consistent in space when the effective averaging time is less. Less averaging time implies less smoothing and leads to noisy measurements of the IVD distribution peaks, as depicted in Fig.7b. As the effective averaging time increases, IVD distribution becomes more regular and denoised to reveal reliable speaker information (Fig.7c and Fig.7d). However, averaging over a much longer period of time leads to redundant spatial images of moving speakers, resulting in inaccurate speaker tracking, and possibly multiple peak detections for the same speaker (Fig.7e and Fig.7f).

5.6. Peak Detection and Tracking

The average IVD distributions of two and four stationary speakers over 2.5s of activity have been depicted in Fig.8a and Fig.8b respectively. As shown in Fig.8b, with lower IVD distribution peaks and a raised floor, the prominence of speaker IVD distributions decrease with an increase in the number of speakers. In a silent environment the IVDs are uniformly distributed in space. As one or two speakers become active, the IVD distribution floor is low and the IVD distribution peaks are prominent. As more speakers start to contribute to the mixture, reflections and reverberations in the surrounding environment escalate. Also, the speakers start to overlap in the T-F domain due to DFT constraints, resulting in an increased violation of W-disjoint orthogonality. Due to less significant IVD distribution of moving

Figure 9: RoIs as ideal circular binary filters for source peaks at 25, 100, 160 and 230 degrees.

speakers with an increase in their number, the process of speaker detection, tracking and separation becomes a challenging problem to solve over short time durations.

In order to deal with a greater number of moving speakers, the proposed BMSS system has to adapt over short time durations to have reliable speaker estimates. Several steps are carried out to detect a potential speaker n , while having the least possible number of false detections. The pseudocode for potential speaker detection and tracking is given in Algorithm 1, where $\hat{K}(t)$ is the average IVD distribution floor, $\hat{c}(t)$ the time varying threshold for IVD peak detection, $N(t)$ the number of speakers, p, q and Q are empirical constants, $peaks\{\cdot\}$ the peak search function, $\theta_n(t)$ the estimated peak locations, $\mathcal{O}_n(t)$ the Region of Interest (RoI), $\epsilon_n(t)$ the location expectation, and $f(\cdot)$ a binary decision function.

The $\hat{K}(t)$ is approximated by averaging the determined troughs in the IVD distribution. Fig.8 shows a decrease in the average peak level of IVD distribution with an increase in the number of speakers. Based on this observation it is assumed that the speaker IVD distribution peaks degrade linearly with the increase in the number of speakers, and the threshold for peak detection, $\hat{c}(t)$, is estimated as in Algorithm 1 Line 5. As a result the effective threshold for peak detection is dynamically lowered in order to perform reliable speaker detection with an increasing number of speakers. If the threshold for peak detection is not adaptive, e.g. with a low static threshold, noisy peaks are also detected as potential speakers. On the other hand, a high static threshold will lead to unreliable and slow detection of the speakers.

After the IVD distribution peaks have been identified for potential speaker activity at a given time frame, the IVD distribution may change due to a

Algorithm 1 Detecting and Tracking Active Moving Speakers

```

1: Input:  $H(IVD; (\omega, t))$ 
2: Output: binary decision  $f(\epsilon_n(t))$ , and respective location  $\theta_n(t)$ 
3: Initialization:  $N(t) = 0, Q = 7.2, p = 0.0077, q = 0.0625, \gamma = 25$ 
4: estimate IVD distribution floor  $\rightarrow \hat{K}(t)$ 
5: estimate peak threshold  $\hat{c}(t) = -N(t)/Q + \hat{K}(t) + 1$ 
6: estimate peak location  $\theta_n(t) = peaks\{H(IVD; (\omega, t)), \hat{c}(t)\}$ 
7: if estimated peak  $H(IVD; \theta_n(t)) > \hat{c}(t)$  then
8:   if  $\theta_n(t) \in \mathcal{O}_n(t)$  where  $f(\theta_n(t)) = 1$  then
9:     update:  $\epsilon_n(t) = \min(\epsilon_n(t-1) + q, 1)$ 
10:  else
11:    declare a new  $\mathcal{O}_n(t)$ ,  $\epsilon_n(t) = q$ 
12:  end if
13:  update:  $\mathcal{O}_n(t) = [\theta_n(t) - \gamma, \theta_n(t) + \gamma]$ 
14: else
15:  update:  $\epsilon_n(t) = \max(\epsilon_n(t-1) - p, 0)$ ,
16:  delete potential speaker if  $\epsilon_n(t) = 0$ ;
17: end if
18: if  $\epsilon_n(t) > 0.5$  then
19:    $f(\epsilon_n(t)) = 1$ ;  $\rightarrow$  speaker detected
20:   update:  $N(t) = N(t-1) + 1$ 
21: else
22:    $f(\epsilon_n(t)) = 0$ ;  $\rightarrow$  speaker inactive
23:   update:  $N(t) = N(t-1) - 1$ 
24: end if

```

change in speaker content or speaker movement in the time frames to follow. To keep track of such time-varying peaks, time-varying RoIs and $f(\mathcal{O}_n(t))$ are proposed to link collocating peaks. RoI is defined as an ideal circular binary filter spanning γ degrees on either side to allow for interference-free movement and satisfactory source separation. Each RoI is updated using $\theta_n(t)$, as given in Algorithm 1 Line 13. We chose γ equal to 25 degrees to allow the sources to be located sufficiently away from each other in order to give satisfactory separation results. For some instantaneous speaker locations, the RoIs are depicted in Fig. 9. Hypothetically, the chosen γ value leads to a possible simultaneous definition of a maximum of $Q = 360/(25 \times 2) = 7.2$ non-overlapping RoIs. If the peak relocates in a particular RoI, it is linked to

Figure 10: Location expectation of four moving speakers with different waiting times. (a) and (c) Speaker 3 in a short period of time (35 frames). (b) and (d) Complete surrounding space in 2200 frames. The red eclipses in (b) and (d) show false detections.

the previously detected peak of the same region. If any of the found $\theta_n(t)$ is closer than γ degrees, the peak with the higher IVD distribution is considered while the other is ignored. Hence, independent of other speakers, a potential speaker can be spatially isolated and tracked based on the IVD distribution representation.

5.7. Location Expectation and Speaker Detection

To effectively perform reliable speaker detection and tracking while dealing with start-stop nature of speech and noisy IVD distribution peaks over short time durations, an expectation-based algorithm is proposed. After detecting the IVD distribution peaks as potential speakers and defining RoIs, the peaks are checked for consistency in the time frames to follow. For each

detected peak a location expectation, $\epsilon_n(t)$, is assigned. The value of $\epsilon_n(t)$ is independent for each RoI, and updated as given in Algorithm 1 Lines 7-16. Speaker detection and separation in a blind manner is directly dependent on the proposed location expectation. If location expectation is greater than 0.5, the corresponding RoI is labeled as belonging to an active speaker. The value of q is set high to reflect swift speaker detection. However, a trade-off exists between swift and reliable detection of the speaker activity. Fig.10a and Fig.10c show the location expectations of an isolated speaker in a short period of time, and Fig.10b and Fig.10d show the respective location expectations for the complete surrounding space. The comparison of the two waiting times considering false speaker detections can be observed in Fig.10b and 10d (indicated by the red eclipses). The number of false speaker detections reduce while trading off time delay (Fig.10d). In current work, the increment in location expectation is set to reflect 185.8ms of speaker activity in accordance with the IIR filtering coefficient. With 2048 window length and 50% overlap, q measures out to be 0.0625. On the other hand, p is set low because of the start-stop nature of speech. For example, a person may pause in between words, and for a longer time in between utterances. 1.5s was empirically found to be a sufficient length of waiting time for the speaker to be declared as completely inactive. It is reflected by the value of p equal to 0.0077.

5.8. Spatial Filtering and Time Domain Representation

The unmixing spatial filters are formed for values of location expectation greater than 0.5. Source separation is achieved by soft masking T-F units to filter out IVD measurements with the help of source specific unmixing spatial filters. The n th source estimate, $\hat{s}_n(\omega, t)$, is obtained as:

$$\hat{s}_n(\omega, t) = W(\omega, t) \times f(\theta; \mu(n, t), \sigma_n), \quad (17)$$

where $f(\theta; \mu(n, t), \sigma_n)$ represent the von-Mises unmixing spatial filters having time-variant locations $\mu(n, t)$ and static beam-width σ_n .

At the final stage, after T-F units of different speakers have been filtered, IDFT and synthesis is carried out to get the separated sources in time domain.

6. BMSS Evaluation and Discussions

Performance evaluation of the proposed BMSS system is done for both stationary and moving speaker scenarios, with two, three and four speaker mixtures.

6.1. Obtaining the Mixtures

Physically stationary and moving speaker recordings were made with human subjects in an ITU-R BS1116 standard listening room, with a reverberation time of 0.32s. The stationary speaker locations and moving speaker regions are depicted in Fig.11a and Fig.11b respectively. Speakers were asked to read pre-defined texts while being stationary or walking in the allocated region of space to avoid co-location. The recording time of all the texts was in between 25s and 35s. The speakers were directed to pause halfway into the recordings. The pause implies no speaker activity in order to test the proposed BMSS system with time-variant mixing parameters.

The sampling rate of the recordings was 44.1kHz, made with the Core Sound TetraMic microphone system (Core-sound.com, 2017). The microphone array was set up 1.5m high above the floor and placed near the center of the room, as depicted in Fig.11. The acoustic axis of the speakers was towards the microphone array in the stationary speaker scenario, while sideways in the moving speaker scenario. At no point in time did the speakers deliver their recordings while directly facing the microphone array. The distance of the speakers to the microphone array was fixed at 1.2m at all times. A-Format data were recorded for individual stationary and moving speaker scenarios. There were three male speakers and one female speaker. Speaker mixtures were later obtained by summing the reverberant recordings of individual speakers in the desired combinations.

6.2. Performance Measures

The BSS Eval (Vincent et al., 2006) and PEASS (Emiya et al., 2011) MATLAB toolboxes were used to evaluate the proposed system performance. In BSS Eval, the signal to distortion ratio (SDR) and the signal to interference ratio (SIR) quantify the separated output signals in terms of overall sound quality and interference, respectively. The estimated i th source, $\hat{s}_i$, of the true source, s_i , is decomposed as:

$$\hat{s}_i = s_i + e_i + e_a + e_n, \quad (18)$$

Figure 11: Experimental setup. (a) Eight stationary source locations labelled as $\bar{a}$ to $\bar{h}$. (b) Eight regions of speaker movement labelled as a to h . A maximum 60 degree interval of speaker movement is expected in each region for the designed setup.

where s_t is the modified version of the true source signal by an allowed distortion, and e_i , e_a and e_n represent the interference, artefacts and noise error terms, respectively. With the help of such breakdown SDR is calculated in (Vincent et al., 2006) as:

$$SDR = 10 \log_{10} [s_t^2 / (e_i + e_n + e_a)^2]. \quad (19)$$

At low angular intervals, the SIR measure gives reliable information by independently quantifying interference from other active speakers. SIR is given in (Vincent et al., 2006) as:

$$SIR = 10 \log_{10} (s_t^2 / e_i^2). \quad (20)$$

A set of perceptually motivated objective measures that aim to predict the subjective scores with accuracy have been proposed in (Emiya et al., 2011). Collectively known as the *Perceptual Evaluation methods for Audio Source Separation* (PEASS), the measures provide a high correlation with the subjectively compiled results. Four quality scores have been suggested in PEASS, which include the artefact related perceptual score (APS), interference related perceptual score (IPS), target related perceptual score (TPS) and the overall perceptual score (OPS). APS measures the artefacts in the target speaker signal, IPS measures the amount of interference coming in

from other speakers, TPS measures the closeness of the separated signal with the clean reference signal and the OPS gives an overall quality score. PEASS scores are on a scale between 0 and 100. Cardioid signals directed towards the target speaker region have been chosen as the reference for carrying out objective evaluation.

In the evaluation, we consider both moving and stationary source cases and test the performance of our proposed BMSS system in difference source movement combinations. The algorithms by Gunel et al. (2008) and Mandel et al. (2010) are used as the baseline for comparison of separation performance. To compare with the baseline fairly, only separation results for stationary sources are used as the baseline is not designed for moving source separation. The separation results including the comparison with the baseline algorithms are presented in Section 6.3.

6.3. Experimental Results

In this section, we first test the detection, tracking and separation performance of the proposed BMSS system with time varying source numbers and source locations. We then present the separation results for comparison with the baseline algorithms.

6.3.1. Detection, Tracking and Separation in BMSS

The recordings were summed to acquire *bd*, *be*, *bf*, *bg* and *bh* two moving source mixtures for different regions labelled in Fig.11b. As examples, the outcomes of the peak detection and location expectation for *bh* and *be* mixtures are depicted in Fig.12a and Fig.12c, while their respective time-variant number of sources are shown in Fig.12b and Fig.12d. The true activity of the speakers is shown in comparison to activity determined by the proposed BMSS system. In Fig.12a and Fig.12c, the speakers are silent at the beginning and are detected once they become active. The dependency on both the time and location are unknown, making the system completely blind to the possible varying mixing parameters. As a speaker becomes active, the peak detection and location expectation algorithm assigns a RoI, and subsequently with consistent peak detection the potential speaker is labelled as an active speaker.

The tracking figures depict the zigzag movement carried out by speakers with respect to the microphone array. Two distinctive tracking streams in Fig.12a and Fig.12c, and a fall in the source number can be observed towards the middle frames in Fig.12b and Fig.12d. The speakers were directed to

Figure 12: Tracking moving speech sources by BMSS system. (a), (c) and (e) Source tracking in regions *bh*, *be* and *aceg* respectively. (b), (d) and (f) Number of sources detected in regions *bh*, *be* and *aceg* respectively. Separate ‘tracks’ of speaker movements can be observed in space in (a), (c) and (e). Near the time frames in the middle (approx. 800 - 1200), the speakers take a pause and start speaking again.

Figure 13: Magnitude spectrum of a separation example for two moving sources in region *be*. (a), (c) and (e) Magnitude spectrum of the B-format mixture signals W, X and Y, respectively. (b) Magnitude spectrum of the original speaker *b*. (d) Magnitude spectrum of the separated speaker *b*. (f) Magnitude spectrum of the separated speaker *e*.

pause after reading the first half of the text and possibly change to a new location within the assigned region based on their own preference. Therefore the dependency of the speaker momentum on time and, on some occasions, space, is broken. As such, the speaker is detected again with different mixing parameters. The speakers can also be seen to abate in location expectation as they become quiet.

Fig.13 shows the examples of magnitude spectrum for two moving sources separation in region *be*. It includes the spectrum of B-format mixture signals W, X, and Y in Fig.13a, Fig.13c and Fig.13e respectively, the original source in region *b* in Fig. 13b, the separated speaker in region *b* in Fig. 13d and the second separated speaker in region *d* in Fig.13f.

The BSS Eval and PEASS evaluation results of separating two moving speakers are shown in Table 1. As the regions are further apart there is a greater spatial discrimination, and as a result less interference for the target speaker from the other speaker. Interestingly, on some occasions the OPS does not show similar trends in results as the SDR, leading to a difference in the perceptual and physical aspects of the separated output signal. For example, for the mixture *bf* the BSS Eval measures comparatively good results but the OPS is on the lower end. It is important to note that the separation performance is dependent on the relative energy and DRR of the speakers as well as their angular interval. As shown in Section 4.2, with increasing energy and DRR the impact of a speaker on T-F components becomes greater, leading to a better representation of that speaker. A speaker which is comparatively low in energy has a less significant impact on the T-F units, and as such, its separated output is of poor quality.

For four moving sources, the BMSS system was evaluated with *aceg* and *bdfh* speaker mixtures. As an example, the results of the speaker tracking and detection for *aceg* mixtures are depicted in Fig.12e and Fig.12f, respectively. The results of the four moving speakers' separation are given in Table 2. With an increased number of speakers the performance degrades, partially due to the added reverberations of more speakers. Also, added speakers lead to more overlapping T-F components, rendering the speaker IVD distribution representation more noisy and unpredictable over short time durations. As such, some T-F components are misdirected and may get registered as interference in the target speaker output. It is reflected by the lower scores of BSS Eval and PEASS measures in Table 2 than the scores in Table 1.

Table 1: Separation performance for two moving sources using BSS Eval (dB) and PEASS

	<i>bd</i>	<i>be</i>	<i>bf</i>	<i>bg</i>	<i>bh</i>	<i>Mean</i>
SDR	7.17	6.92	7.11	7.99	7.95	7.43
SIR	21.77	21.65	23.99	24.83	22.45	22.94
OPS	45.06	37.71	37.28	44.83	43.66	41.71
TPS	64.32	60.55	58.66	60.51	62.41	61.29
IPS	73.24	75.51	75.76	72.93	74.45	74.34
APS	43.23	35.35	35.68	44.92	41.02	40.04

Table 2: Separation performance four moving sources using BSS Eval (dB) and PEASS

	<i>aceg</i>	<i>bdfh</i>	<i>Mean</i>
SDR	4.44	4.59	4.52
SIR	15.50	15.87	15.69
OPS	32.17	33.93	33.05
TPS	57.27	53.92	55.60
IPS	71.20	69.57	70.39
APS	31.42	37.59	34.51

6.3.2. Comparison with the Baseline

The separation performance of the proposed BMSS system is compared with two other source separation algorithms, the algorithm of Gunel et al. (2008), which we refer to as Gunel, and the algorithm of Mandel et al. (2010), which we refer to as Mandel. We implemented the Gunel algorithm, and downloaded the Matlab code of Mandel algorithm from the website (Mandel and Weiss, 2009).

Comparison with the Gunel algorithm is straightforward as it takes A-format signals as input directly. In order to generate the binaural mixtures for the Mandel algorithm, the individual speaker’s recordings are first convolved with the binaural room impulse response (BRIR) measured with a Cortex Instruments Mk2 dummy head in an anechoic room. The binaural mixtures are then obtained by summing the convolved individual speakers at the desired angular displacements and regions.

The overall performance of separation for two and three sound source mixtures is also evaluated with BSS Eval and PEASS, and the average results

Table 3: Overall separation performance for two and three sound source mixtures using BSS Eval (dB) and PEASS.

Algorithm	2 Sources			3 Sources		
	Proposed	Gunel	Mandel	Proposed	Gunel	Mandel
SDR	11.94	10.55	11.65	7.55	6.23	6.95
SIR	24.95	24.18	23.48	18.62	17.37	12.38
OPS	68.58	59.45	50.68	44.92	37.16	38.41
TPS	74.58	67.27	66.80	63.75	69.76	45.71
IPS	72.14	75.13	64.69	72.34	69.95	56.12
APS	60.38	43.47	46.56	45.51	33.41	40.38

of sound sources located at different angular displacements are summarised in Table 3. For the case of two sound sources, they are placed in five different location combinations, namely $\bar{b}\bar{d}$, $\bar{b}\bar{e}$, $\bar{b}\bar{f}$, $\bar{b}\bar{g}$, and $\bar{b}\bar{h}$ as shown in Fig.11a, and their angular distances are 90, 135, 180, 225, 270 degrees respectively. For three sound sources, their location combinations include $\bar{a}\bar{c}\bar{d}$, $\bar{a}\bar{c}\bar{e}$, and $\bar{a}\bar{d}\bar{e}$, which corresponds to the sound sources being placed at 0-90-135, 0-90-180, 0-135-180 degrees locations, respectively.

From Table 3, it is clear that, for two sources, the proposed system outperforms over the baseline algorithms in all evaluated parameters, except the IPS score. Comparing with the Gunel algorithm, the proposed system gives about 1.5 dB improvement in SDR and 0.8 dB in SIR. It also shows an improvement of about 0.3 dB in SDR and 1.5 dB in SIR over the Mandel algorithm. Significant improvement, about 9 points and 18 points in the overall perceptual score OPS, has been achieved over the two baseline algorithms. This means the separated sound from the proposed system will have lesser distortion and interference from other sound sources and much reduced artefacts than the baseline algorithm.

The overall better performance of the proposed BMSS system over the baseline algorithms is also evidenced in the three sources separation results shown in Table 3. Note that for three sources separation, the SDR of the proposed system is about 1.3 dB better than the Gunel algorithm and 0.6 dB better than the Mandel algorithm, the SIR has an improvement of about 1.3 dB and 5.7 dB over the baselines, and the overall perceptual score OPS outperforms the two baselines by about 7.8 points and 6.5 points, respectively. Overall, the performance improvement of the proposed BMSS system

for three sources separation is generally in line with our observations for two sources separation. From the table, it is clear that the performance, for all evaluated parameters, of all the algorithms decreases as the source numbers are increased from two to three in the experiment. This suggests that all the algorithms rely on the spectral cues for separation. More sources means more overlapping T-F components and higher interference in the separated target from other sound sources.

Fig.14 shows a comparison of the magnitude spectrum of two sources separation by different algorithms. The two sources, *bf*, are located with an angular displacement of 180 degrees. The magnitude spectrum of the separated signals by the baseline Gunel, baseline Mandel, and the proposed BMSS method are shown in Fig.14d, Fig.14e and Fig.14f.

7. Conclusion

In this paper a novel IVD-based algorithm is proposed to blindly cater for the time-variant mixing conditions in acoustic moving sources separation. Denoising of the IVD distribution is carried out using the von-Mises modelling in space and IIR filtering in time to extract reliable speaker estimates over short time durations. It was found that in order to perform speaker detection and tracking with reliability, system adaptation to changing mixing conditions is critical.

The proposed system adapts the threshold for the IVD distribution peak detection in order to reliably detect new speakers. It was established that the speaker's energy and DRR dictate the speaker detection rate and the separated output quality. A low energy, low DRR speaker has noisy IVD distribution and poor separated output quality in the presence of comparatively more energetic speakers. However, during natural speech activity, drastic energy variations occur over short time durations, and the movement of the speakers adds to the already challenging problem. In such scenarios applying IIR filtering on IVD distribution and using algorithms of adaptive thresholds, RoI and location expectation all help in detecting and tracking the true speaker peaks reliably.

Objective evaluation of the proposed BMSS system was carried out with BSS Eval and PEASS toolboxes. It showed that the system outperformed the baseline Gunel algorithm and also Mandel's algorithm.

Acknowledgments

This work was supported by MulSys Limited through the European Commission (Grant Number 287896). We would like to thank Dr Wenwu Wang for help in evaluation. We would also like to thank the anonymous reviewers for their valuable comments in improving this paper.

References

References

- Addison, W., Roberts, S., 2006. Blind source separation with non-stationary mixing using wavelets, in: Proc. Intl. Conf. Independent Component Analysis and Blind Source Separation.
- Alinaghi, A., Jackson, P.J., Liu, Q., Wang, W., 2014. Joint mixing vector and binaural model based stereo source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 1434–1448. doi:10.1109/TASLP.2014.2320637.
- Anemüller, J., Gramssý, T., 1999. On-line blind separation of moving sound sources, in: Proc. ICA’99, Aussois, France.
- Araki, S., Makino, S., Hinamoto, Y., Mukai, R., Nishikawa, T., Saruwatari, H., 2003a. Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures. *EURASIP J. Adv. Signal Process.* 2003, 1157–1166.
- Araki, S., Mukai, R., Makino, S., Nishikawa, T., Saruwatari, H., 2003b. The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Speech Audio Process.* 11, 109–116.
- Bao, G., Ye, Z., Xu, X., Zhou, Y., 2013. A compressed sensing approach to blind separation of speech mixture based on a two-layer sparsity model. *IEEE Trans. Audio, Speech, Language Process.* 21, 899–906.
- Benjamin, E., Chen, T., 2005. The native b-format microphone, in: Audio Engineering Society Convention 119.

- Bree, H.E., Druyvesteyn, W., Berenschot, E., Elwenspoek, M., 1999. Three-dimensional sound intensity measurements using microflown particle velocity sensors, in: Proc. MEMS'99 Twelfth IEEE International Conference on Micro Electro Mechanical Systems, pp. 124–129.
- Cardoso, J.F., 1998. Blind signal separation: statistical principles. Proc. IEEE 86, 2009–2025.
- Cherry, E.C., 1953. Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. 25, 975–979.
- Chong, N., Shanhung, W., Nordholm, S., Murray, I., 2014a. Multiple sound source tracking and identification via degenerate unmixing estimation technique and cardinality balanced multi-target multi-bernoulli filter (DUET-CBMeMber) with track management, in: Proc. APSIPA, pp. 1–5.
- Chong, N., Wong, S., Vo, B.T., Sven, N., Murray, I., 2014b. Multiple moving speaker tracking via degenerate unmixing estimation technique and cardinality balanced multi-target multi-bernoulli filter (DUET-CBMeMber), in: Proc. IEEE 9th ISSNIP, pp. 1–6.
- Comon, P., Jutten, C., 2010. Handbook of Blind Source Separation: Independent component analysis and applications. 1st ed., Academic press.
- Core-sound.com, 2017. Introduction: Tetramic single point stereo & surround sound microphone URL: <http://www.core-sound.com/TetraMic/1.php>.
- Douglas, S.C., Gupta, M., Sawada, H., Makino, S., 2007. Spatio-temporal FastICA algorithms for the blind separation of convolutive mixtures. IEEE Trans. Audio, Speech, Language Process. 15, 1511–1520.
- Emiya, V., Vincent, E., Harlander, N., Hohmann, V., 2011. Subjective and objective quality assessment of audio source separation. IEEE Trans. Audio, Speech, Language Process. 19, 2046–2057.
- Griffiths, L.J., Jim, C.W., 1982. An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antennas Propag. 30, 27–34.
- Günel, B., Hacıhabiboğlu, H., Kondo, A.M., 2007. Wavelet packet based analysis of sound fields in rooms using coincident microphone arrays. Applied Acoustics 68, 778–796.

- Gunel, B., Hacıhabiboglu, H., Kondo, A.M., 2008. Acoustic source separation of convolutive mixtures based on intensity vector statistics. *IEEE Trans. Audio, Speech, Language Process.* 16, 748–756.
- Heinzel, G., Rüdiger, A., Schilling, R., 2002. Spectrum and spectral density estimation by the discrete fourier transform (dft), including a comprehensive list of window functions and some new at-top windows .
- Huang, P., Kim, M., Hasegawa-Johnson, M., Smaragdis, P., 2014. Deep learning for monaural speech separation, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1562–1566.
- Hyvarinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* 10, 626–634.
- Khan, M., Naqvi, S., Ur-Rehman, A., Wang, W., Chambers, J., 2013. Video-aided model-based source separation in real reverberant rooms. *IEEE Trans. Audio, Speech, Language Process.* 21, 1900–1912.
- Koldovsky, Z., Malek, J., Tichavsky, P., Nesta, F., 2013. Semi-blind noise extraction using partially known position of the target source. *IEEE Trans. Audio, Speech, Language Process.* 21, 2029–2041.
- Koutvas, A., Dermatas, E., Kokkinakis, G., 2000. Blind speech separation of moving speakers in real reverberant environments, in: *Proc. IEEE ICASSP’00*, pp. 1133–1136.
- Mandel, M., Weiss, R., 2009. Model-based em source separation and localization. URL: <https://github.com/mim/messl>.
- Mandel, M.I., Weiss, R.J., Ellis, D.P.W., 2010. Model-based expectation-maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing* 18, 382–394. doi:10.1109/TASL.2009.2029711.
- Mardia, K.V., Jupp, P.E., 2009, pp.36–45. *Directional statistics*. John Wiley & Sons.
- Merimaa, J., Pulkki, V., 2005. Spatial impulse response rendering i: Analysis and synthesis. *J. Audio Eng. Soc.* 53, 1115–1127.

- Mitianoudis, N., 2012. A generalized directional laplacian distribution: Estimation, mixture models and audio source separation. *IEEE/ACM Trans. Audio, Speech, Language Process.* 20, 2397–2408.
- Mitianoudis, N., Davies, M.E., 2004. Audio source separation: Solutions and problems. *Int. J. Adapt. Control Signal Process.* 18, 299–314.
- Naqvi, S., Wang, W., Khan, M., Barnard, M., Chambers, J., 2012. Multimodal (audio-visual) source separation exploiting multi-speaker tracking, robust beamforming and time-frequency masking. *IET Signal Processing* 6, 466–477.
- Naqvi, S., Yu, M., Chambers, J., et al., 2010. A multimodal approach to blind source separation of moving sources. *IEEE J. Sel. Topics Signal Process.* 4, 895–910.
- Nugraha, A.A., Liutkus, A., Vincent, E., 2016a. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 1652–1664.
- Nugraha, A.A., Liutkus, A., Vincent, E., 2016b. Multichannel music separation with deep neural networks, in: 2016 24th European Signal Processing Conference (EUSIPCO), pp. 1748–1752.
- Olson, H., 1957. *Acoustical Engineering*. 3rd ed., Van Nostrand.
- Parra, L.C., Alvino, C.V., 2002. Geometric source separation: Merging convolutive source separation with geometric beamforming. *IEEE Speech Audio Process.* 10, 352–362.
- Plinge, A., Hennecke, M.H., Fink, G.A., 2012. Reverberation-robust online multi-speaker tracking by using a microphone array and CASA processing, in: *Proc. IWAENC*, pp. 1–4.
- Prieto, R.E., Jinachitra, P., 2005. Blind source separation for time-variant mixing systems using piecewise linear approximations, in: *Proc. ICASSP’05*.
- Princen, J.P., Bradley, A.B., 1986. Analysis/synthesis filter bank design based on time domain aliasing cancellation. *IEEE Trans. Acoust., Speech, Signal Process.* 34, 1153–1161.

- Pulkki, V., 2007. Spatial sound reproduction with directional audio coding. *J. AUDIO ENG.* 55, 503–516.
- Rickard, S., 2007. The DUET blind source separation algorithm, in: *Blind Speech Separation*. Springer Netherlands, pp. 217–241.
- Rickard, S., Yilmaz, O., 2002. On the approximate w-disjoint orthogonality of speech, in: *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I–529–I–532. doi:10.1109/ICASSP.2002.5743771.
- Sarmiento, A., Durán-Díaz, I., Cichocki, A., Cruces, S., 2015. A contrast function based on generalized divergences for solving the permutation problem in convolved speech mixtures. *IEEE/ACM Trans. Audio, Speech, Language Process.* 23, 1713–1726.
- Sawada, H., Araki, S., Makino, S., 2011. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Trans. Audio, Speech, Language Process.* 19, 516–527.
- Shindo, H., Hirai, Y., 2002. Blind source separation by a geometrical method, in: *Proc. IEEE IJCNN’02*, pp. 1109–1114.
- Smaragdis, P., 1998. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing* 22, 21–34.
- Souden, M., Araki, S., Kinoshita, K., Nakatani, T., Sawada, H., 2013. A multichannel MMSE-based framework for speech source separation and noise reduction. *IEEE Trans. Audio, Speech, Language Process.* 21, 1913–1928.
- Sturmel, N., Daudet, L., 2013. Informed source separation using iterative reconstruction. *IEEE Trans. Audio, Speech, Language Process.* 21, 178–185.
- Traa, J., Smaragdis, P., 2013. Blind multi-channel source separation by circular-linear statistical modelling of phase differences, in: *Proc. ICASSP’13, Vancouver, BC, Canada*. pp. 4320–4324.
- Traa, J., Smaragdis, P., 2014. Multichannel source separation and tracking with RANSAC and directional statistics. *IEEE/ACM Trans. Audio, Speech, Language Process.* 22, 2233–2243.

- Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N., Mitsufuji, Y., 2017. Improving music source separation based on deep neural networks through data augmentation and network blending, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 261–265.
- Van Der Veen, A.J., 1998. Algebraic methods for deterministic blind beamforming. *Proc. IEEE* 86, 1987–2008.
- Vincent, E., Gribonval, R., Févotte, C., 2006. Performance measurement in blind audio source separation. *IEEE Trans. Audio, Speech, Language Process.* 14, 1462–1469.
- Wang, D., Chen, J., 2018. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 1702–1726.
- Yamamoto, S., Nakadai, K., Nakano, M., Tsujino, H., Valin, J.M., Komatani, K., Ogata, T., Okuno, H., 2006. Real-time robot audition system that recognizes simultaneous speech in the real world, in: *Proc. IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, pp. 5333–5338.
- Yamashita, J., Tatsuta, S., Hirai, Y., 2004. Estimation of propagation delays using orientation histograms for anechoic blind source separation, in: *Proc. IEEE IJCNN'04*, pp. 2175–2180.
- Yilmaz, O., Rickard, S., 2004. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Signal Process* 52, 1830–1847.
- Yoshioka, T., Nakatani, T., Miyoshi, M., Okuno, H., 2011. Blind separation and dereverberation of speech mixtures by joint optimization. *IEEE Trans. Audio, Speech, Language Process.* 19, 69–84.
- Yu, Y., Wang, W., Han, P., 2016. Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks. *EURASIP Journal on Audio, Speech, and Music Processing* 2016, 7. URL: <https://doi.org/10.1186/s13636-016-0085-x>, doi:10.1186/s13636-016-0085-x.

Zhao, W., Shen, Y., Xu, P., Wang, J., Yuan, Z., Wei, Y., Jian, W., Li, H., 2014. A new efficient method for permutation and scaling ambiguity of blind source separation signal blocks, in: Proc. Fifth International Conference on Intelligent Control and Information Processing, pp. 23–28.

Figure 14: Comparison of magnitude spectrum of two sources separation by different algorithms. (a) Magnitude spectrum of the omni-directional signal, W , for the two sources at bf . (b) Magnitude spectrum of original speaker b . (c) Magnitude spectrum of the binaural mixture of two speakers at location bf . (d) Magnitude spectrum of separated speaker b by baseline Gunel. (e) Magnitude spectrum of separated speaker b by baseline Mandel. (f) Magnitude spectrum of separated speaker b by the proposed BMSS system.