A Survey of Blind Source Separation

John Yeary

Abstract

Blind Source Separation (BSS) is a robust technique that involves separating multiple non stationary signals without having any prior information of signal properties or mixture. A survey of various techniques applied to BSS ranging from numerical methods based around matrix factorization and neural network, deep learning based methods. Several of these methods are described, and various strengths and weaknesses of these approaches are compared.

1. Introduction

This survey provides an extended review of various methods dedicated towards solving BSS, a signal processing technique aimed at isolating and reconstructing individual sources within a received mixed signal. This area has received a lot of recent attention as the problem is extensible to many domains of signal processing, ranging from ultrasound imaging [16] to RADAR [7]. However this paper will focus on the common domain for solving this problem, processing audio signals.

The surveyed techniques will provide a broad overview of different approaches to resolve this common problem, often geared towards different relevant cases of BSS. While there are a diverse range of approaches to solve BSS, the most simplistic is time-frequency analysis (TFA). This analysis combines frequency domain information with timing of the signal, vital for the non-stationary nature of audio sources. While the short time frequency transform (STFT) is a common approach, continuous wavelet transform (CWT) and Wigner-Ville Distribution (WVD) [9] can also be applied to inspect the time and frequency dependence of the signals mixed together. In the most simplistic case, signals of different frequencies can be directly extracted through this method. Other methods attempt to selectively mask time-frequency bins to identify the contributions from different sources based on the time frequency structure of the received signal.[17]

In general, TFA is used as a building block for many other BSS methods.Typically BSS approaches apply some preprocessing to de-noise and de-correlate received channel signals. A standard formulation of the BSS problem relies on using frequency domain information, and given the non-stationary nature of speech or music signals, this requires some time representation as well. After applying STFT or another TFA technique, BSS approaches fall into two broad categories: statistical based and deep learning. Finally, TFA is used again to convert the now isolated signals to the time domain.

One of the most recognized statistical approaches is Independent Component Analysis (ICA). ICA, and other statistical approaches inheriting from ICA, assume that the source signals to estimate are statistically independent and non-Gaussian. [4] This technique typically maximizes non-Gaussianity, or attempts to minimize the mutual information between estimated sources. Relaxing the restriction for true independence, second order statistical methods, and joint diagonalization approaches focus on the decorrelation of estimated sources [6]. This approach is often a better model of the BSS problem at hand and is well suited to instances where reverberations are present or when the observed signals are not necessarily independent (such as RADAR applications).[10] Another general statistical approach is nonnegative matrix factorization (NMF). This approach further relaxes the decorrelated assumption and instead uses signal nonnegativity to estimate the original sources, a good assumption when dealing with spectral data. Finally, one can assume source sparsity. All of these approaches do reasonably well under well conditioned problems (overdetermined or determined systems), but all suffer from the permutation problem, where the output of the BSS technique is not guaranteed to be in the same order from iteration to iteration.

Deep learning approaches in BSS typically require an enormous amount of data in order to prevent overfitting. Due to the sheer number of variations in environmental factors, it is often difficult to capture every potential scenario within a training dataset. Instead, recent developments seek to either isolate the role of neural networks within the BSS framework, or to utilize generative models to embed statistical representations of various speakers during training tries to map observed data to those representations.[11][8] In order to try and prevent overfitting, researchers have suggested using a knowledge distillation scheme where a fully trained model provides 'hints' to the model under training to try and

promote the model's ability to adapt to new data [8]. In general, deep learning approaches still require a large amount of training data to be effective, and many approaches have began pursuing unsupervised or semi-supervised methods. While statistical methods must assume a linear mixing relationship between sources, deep learning based methods do not require any assumptions of the source behavior, making them ideal if the challenges of computational resources and generalizations are solved.[1]

Further enhancements are often made to these algorithms by introducing spatial dependence and beamforming. Importantly, one technique and set of assumptions can be used to analyze the time-frequency signal, and another technique can be applied to the spatial frequency domain. This provides researchers with ample degrees of freedom to pursue new and innovative ideas.

2. Background

Blind Source Separation refers to a set of algorithms and techniques intended to identify, isolate and enhance N independent sources from an unknown received mixed signal observed at M receivers. BSS is typically formulated as a linear mixture model. Let $\mathbf{X} \in \mathbb{R}^{Mx1}$ vector which represents the signal observed at M sensors. By assuming that the mixture of different sources is linear, i.e. no nonlinear mixing occurs, we can represent $\mathbf{X}$ as:

$$\mathbf{X} = \mathbf{AS} \tag{1}$$

Where $\mathbf{A} \in \mathbb{R}^{MxN}$ matrix known as the mixing matrix, that represents the weights of the N individual sources observed at each M sensor. The source vector, $\mathbf{S} \in \mathbb{R}^{N}x1$ represents the source signals. Immediately, A and S are unknown. Rather than attempt to resolve these two unknowns, it is common to attempt to solve the inverse problem.

Using the X vector, we can instead try to find a de-mixing matrix B such that:

$$Y = BX \approx S \tag{2}$$

That is, by making some assumptions about the mixing process and properties of the individual sources, we can try to find a vector Y that approximates the true S vector.

The assumptions made lead to the different numerical algorithms introduced in Section 1, such as non-Gaussianity for ICA based approaches, jointly diagonal for JD algorithms, or non-negativity for NMF based algorithms.

Joint diagonalization is defined as:

Definition 2.1. Given a set of K symmetric matrices $\mathbf{A}_1, \mathbf{A}_2, \dots \mathbf{A}_K$, the matrices are jointly diagonal if a transformation matrix $\mathbf{V}$ exists such that the transformed matrices $\mathbf{D}_k$ are diagonal:

$$\mathbf{D}_k = \mathbf{V}^T \mathbf{A}_k \mathbf{V}, \quad \forall k = 1, 2, \dots K$$

Often times joint diagonalization is impossible due to noise or other imperfections in the data and is extended to approximately joint diagonal.

Non-Gaussianity, a requirement for ICA, is measured as the deviation of the signal from a Gaussian distribution. This is typically calculated by taking the kurtosis or negentropy of the signal, which are defined as:

Definition 2.2. Kurtosis is a measure of 'tailedness' in a distribution. Mathematically, this is:

$$\text{Kurt}(x) = \frac{\mathbb{E}[(x-\mu)^4]}{(\mathbb{E}[(x-\mu)^2])^2} - 3$$

Where:

- x is the random variable representing the signal
- μ is the mean of variable x

Definition 2.3. Negentropy attempts to quantify how much a distribution deviates from a Gaussian distribution. The term 'Negentropy' relates to the fact that Gaussian distributions have a maximum entropy and are the least non-Gaussian, while non-Gaussian distributions should have less entropy, or negative entropy (see A.2). Mathematically, this is given as:

$$J(p_x) = H(\phi_x) - H(p_x) \tag{3}$$

Where:

- $H(p_x)$ is the differential entropy of a random variable x given by: $H(p_x) = -\int p_x(u) log p_x(u) du$
- ϕ_x is a Gaussian distribution with the same mean and variance as p_x

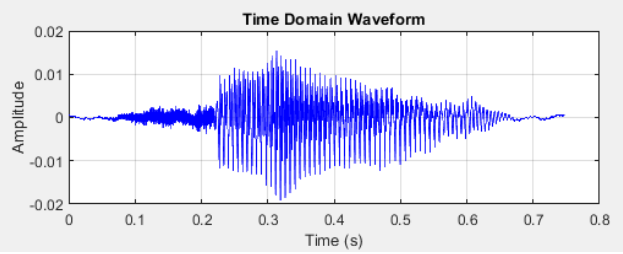
In practice, 3 is difficult to calculate directly, so it is often approximated by more manageable equations.

The assumptions and formulations made here are often used as the cost function when iteratively approaching a solution through Newton's method or another optimization method.

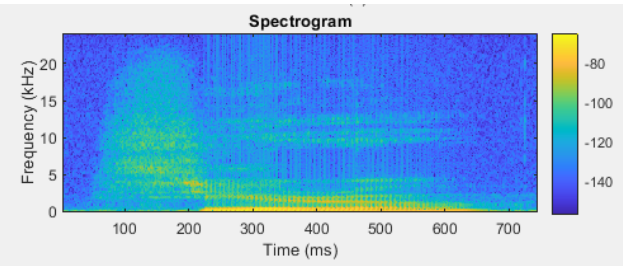
Another vital concept for BSS is time-frequency analysis. The most common approach is to use a short time Fourier transform. Based upon the Fourier transform which decomposes time domain signals into a sum of sinusoidal magnitudes and phases, STFT segments the time domain signal into separate windows and applies the Fourier transform to each window. In discrete time, this is mathematically defined as:

$$X(m, \omega) = \sum_{n=-\infty}^{\infty} x[n] w[n-m] e^{-i\omega n} \tag{4}$$

Where w is a window function. This form creates a spectrogram, where the time and frequency relationship of a signal is easily understood and is a common starting point for many deep learning and statistical methods.



(a) Speech Waveform from AudioMNIST of the word 'one'



(b) Speech Spectrogram from AudioMNIST of the word 'one'

Figure 1. Representations of speech signal

Each column in the spectrogram from figure 1(b) represents a timestep, and each row represents a frequency bin from the Fourier Transform. With this spectrogram representation, algorithms can be applied by window (timestep), or can analyze time-frequency relationships between signals in order to mask specific time frequency bins to attempt to isolate individual signals.

Another relevant concept when discussing sound and acoustic sources is harmonic frequencies. Harmonic frequencies arise from the natural resonant frequencies of a sound source. These frequencies are an integer multiple of a fundamental frequency that is typically referred to as the 'pitch' of a signal, while the harmonics are typically perceived as the 'timbre' of the sound. Formally, these frequencies are given by:

$$f_n = nf_0$$

Where f_0 is the fundamental frequency. In Fourier analysis, a Fourier series representing a sound wave x at time t can be represented as:

$$x(t) = A_0 + \Sigma_{n=1}^{\infty} A_n \cos\left(2\pi n f_0 t + \phi_n\right)$$

Where:

- A_0 is the zero-frequency (DC) magnitude
- A_n is the amplitude of the n-th harmonic
- ϕ_n is the phase shift of the n-th harmonic

Intuitively, the sound of a vibrating source is influenced by the relationship between the amplitude and phases of these harmonics.

Finally, Wiener filters are a vital part of ongoing signal processing research. These are simply a family of filters aimed at minimizing the mean squared error between an estimated signal and a true signal. See appendix A.3 and A.4 for more details.

3. Problem Statement

Research in BSS has been evolving rapidly over the past few years. One of the fundamental approaches to resolving BSS was Independent Component Analysis (ICA). ICA laid the framework described in section 2, where sources were assumed to be linearly mixed according to some mixing matrix $\mathbf{A}$. The initial ideas presented in [5] have been extended to cover multiple sets of assumptions and criterion, leading to such approaches as nonnegative matrix factorization and joint diagonalization. While modern statistical approaches have shown improvement over ICA, there are still several areas of known challenges:

- *Permutation problem*: Many BSS methods operate on individual frequency bins, meaning that the final output might be scrambled between various sources. Independent low rank matrix analysis (ILRMA), nonnegative matrix factorization (NMF) and Independent vector analysis (IVA) attempt to correct this issue by introducing dependencies between frequency bins. Further progress has been made by introducing spatial dependencies through beamforming.
- *Underdetermined*: Cases where the number of sources (N) is greater than the number of receivers (M) is notoriously difficult. Since the system is ill posed, there are potentially infinite different solutions. It is common to make sparsity assumptions in the time, frequency or both domains in order to limit the set of solutions. Time frequency masking is one approach that is aimed at resolving this problem, but struggles when sources overlap in time or frequency. Due to a lack of redundant sampling, noise and other perturbations tend to have a strong effect on source separation performance. Nonlinear approaches such as deep learning are also targeted towards resolving underdetermined systems.
- *Reverberation and Convolutive mixtures*: Most BSS methods assume the linear mixing model. However,

in reality many environments cause reverberant and convolutive mixing that is difficult to resolve using solely statistical techniques. Including spatial features can help, however this issue makes it difficult to recover clean features in time and frequency domains which can impede deep learning approaches as well.

- *Computation*: Many BSS approaches require many matrix inversions to compute separation matrices. Furthermore, BSS methods rely on iterative optimization to approach a global solution to the separation problem. These constraints make it difficult for BSS to be applied online and on platforms with lightweight computing power, such as microprocessors or hearing aides.

These issues are subjects of active research and largely remain open questions.

4. Literature Review

As mentioned in section 3, there are several areas of active research within BSS. One foundational paper in time-frequency masking is presented in [19]. Aimed towards the under-determined problem mentioned in section 3, this paper suggests using binary time frequency masks to separate speech signals from each other. By assuming that the time frequency components of separate signals are not overlapping (a condition the authors call W-disjoint orthogonality), the authors propose an algorithm to identify multiple source signals from two anechoic additive mixtures. Importantly, the authors proved that speech signals are approximately W-disjoint orthogonal and introduce a method to quantify the overlap between signals, a concept that opens the door for more sophisticated time-frequency masking approaches.

Aimed at resolving the permutation problem and reverberant problem, [13] describes an approach which combines spatial dependency with spectral correlations to effectively improve source identification and separation in reverberant conditions. The author proposes using a 2 step process to fix frequency bin alignment between sources. The first step employs direction of arrival (DOA) estimation to fix permutations for frequencies that have high DOA confidence, followed by fixing permutations for highly correlated source envelopes. This is built of the assumption that several highly correlated sources are really the same source, again leveraging an independence assumption. In order to add robustness, the authors include a harmonic correlation measure to help align permutations. They exploit strong correlations between the fundamental and higher ordered harmonics to identify frequencies that likely belong to the same source. If a frequency bin is assigned to a source envelope, the other harmonic frequencies can be added to that source as well. Vitally this paper introduced the idea of using harmonic dependence between ICA resolved frequency bins to correct permutations.

As mentioned previously, deep learning methods have shown great promise in BSS applications. [12] outlines an approach to applying deep neural networks to this problem area by learning the spectrogram of sources. By extending the typical single channel deep learning approach to multiple channels, the authors can utilize spatial information, as well as reduce noise from the signal. The authors propose a delay and sum beamforming step to align signals received at different sources, then pass the resulting signal through the network to attempt to generate a spectrogram representation of each source. Using this spectrogram alongside an expectation-maximization optimization step, the authors propose using a set of Wiener filter parameters to separate each source. The expectation step takes in the estimated power spectral densities (PSD) of the sources from the neural network alongside a spatial covariance matrix to calculate second order moments from the spatial source image derived through multichannel Wiener filtering. The maximization step then updates the spatial covariance estimate, and uses neural networks to update the PSD estimates. The authors suggest using a unique neural network for each iteration in the EM procedure due to changes within the input structure after each iteration. This paper extended the application of deep neural networks to multichannel cases and emphasizes a joint statistical and deep learning based approach to overcome shortcomings of each approach.

[2] addresses issues with convolutive and under-determined BSS. This paper mentions that previous approaches centered around narrowband spatial covariance matrix assumptions, meaning that the convolution was modeled as a complex multiplication by each frequency component in the signal within an STFT window. This is assuming that the convolution filter is the same length or shorter than the window for the STFT. Effectively this assumption reduces the approximation of the spatial covariance matrix used to isolate sources into a rank 1 matrix, meaning that the mixture is assumed to be dominated by a single source. In reality these assumptions do not typically hold since reverberant environments typically have mixing filters whose length is longer than that of the STFT applied to the signal. This paper recommends extending these spatial covariance matrices to be full rank unconstrained in order to properly model spatial dependency in reverberant environments. This new approach allows spatial components to be unrelated a priori, meaning that the model is more flexible and can better fit the observed data. The author describes an expectation-maximization (EM) based algorithm utilizing this new full rank, unconstrained spatial covariance matrix, but acknowledges the increased computational cost incurred by using these full rank matrices.

5. Qualitative Comparison

Recent advancements in BSS have largely revolved around advances in deep learning, leveraging improved processing power to create comprehensive models to separate different types of signals.[1] Generative models such as Variational Auto Encoders (VAE) and Generative Adversarial Networks (GAN) have been applied to BSS problems as these models are naturally good fits for BSS. These models attempt to map input data (x) to some lower dimensional latent space (z), which inherently mirrors the problem presented by BSS.

One such approach is provided by Fast Multichannel Variational Auto Encoder 2 (FastMVAE2), which uses a conditional VAE (CVAE) to model the spectrogram of each speech sample.[8]Building on the multichannel approach from [12], this CVAE is trained using clean speech samples, and a speaker ID is used as the conditioning variable. Aimed at improving the computational efficiency and generalization of the MVAE method, FastMVAE2 introduces the ChimeraACVAE, which is a deep learning model that merges the encoder and classifier networks of its predecessor. This model assumes conditional independence between the two, allowing for parallel processing to expedite computation and decrease error propagation. Further, in order to improve generalizability, FastMVAE2 employs a knowledge distillation training scheme, where a pre-trained CVAE model samples the training data in parallel to the ChimeraACVAE and is used to calculate loss in the ChimeraACVAE that is in training. Importantly, the student model encoder does not depend on the class, while the 'teacher' does have knowledge of the class label for the input. The KL divergence between these distributions is taken and used to calculate loss for the models. See the diagram in A.6 for a depiction.

Through these improvements, FastMVAE2 achieves its goal of improving generalizability and computational performance. The model produces a reduced runtime compared to the original FastMVAE architecture, but is still longer than ILRMA.[8] Further improvements were made to model complexity. FastMVAE2 demonstrates similar separation results to MVAE, but is still limited to determined cases. This approach still struggles in highly reverberant environments. Further work could include dereverberation efforts, which would either introduce more reverberant samples to the training set, or introduce dereverberation techniques to the network.

In contrast to deep learning approaches, there are also active efforts to improve statistical methods. As neural network based BSS requires lots of training data to generalize well, statistical techniques provide a more compact approach. One technique, Harmonic Vector Analysis (HVA) builds upon ideas presented in [13].[17] This approach relies on spectral correlation between sources, leveraging the harmonics of different sources to separate them. Similar to [19], this approach attempts to update a time-frequency mask designed to isolate harmonic sources. The authors first take a threshold of the 'cepstrum' of the signal, which is a Fourier transform of the log amplitude frequency spectrum in order to highlight harmonics in the signal, then iteratively update their filter to isolate these sources. This provides a flexible source separation approach that steps away from statistical independence assumptions, instead relying on harmonic independence. This technique should be well suited for music sources, but will struggle against impulsive, non-harmonic sources like drums.

Also applying a Wiener-like filter, [6] introduces a joint diagonalization to the spatial covariance matrices to estimate and update a multi-channel Wiener filter. This leads to an ICA derived approach to BSS called Full-Rank Spatial Covariance Analysis (FCA) which is aimed towards resolving underdetermined and reverberant cases. An advantage of this approach is again the computational complexity. By assuming joint diagonalizable matrices, these approaches reduces the many matrix inversions typically needed for Wiener filter based BSS.

Computational efficiency is a key highlight of each approach. FastFCA is the only approach that is designed to resolve underdetermined sources and to work well in highly reverberant scenarios. Similarly, there is future work that could extend FastMVAE2 to reverberant cases, potentially incorporating a statistical-based approach.[3] Since HVA exploits harmonic structure it is best suited toward music separation of the three, but might struggle if the instrument's harmonics are similar or overlapped. FastFCA and HVA both assume stationary or approximately stationary distributions which are difficult to enforce in the real world due to room acoustics, speaker movement, and other unknowns. See 4 for a summarization.

6. Numerical Comparison

Code was taken from each of the papers (see A.5). Code for FastFCA and HVA were available in MATLAB, and FastMVAE2 was implemented in Python. To evaluate each method, data from SISEC 2011 was used.[15] Three samples were taken from this dataset: one was two guitars recorded independently, one was two female speakers recorded independently, and another was anechoic speech from two male speakers. The guitar samples and the female speaker samples were recorded via two microphones set up in a room, with speakers distributed within the room. The samples were then played over the speakers one at a time, and the individual responses were recorded through each microphone.

For the male speech samples, the source, anechoic signals were used. These are from a single speaker into a single

channel within an anechoic chamber. These were then multiplied by a random mixing matrix A to generate an additive two channel speech sample. (see table 6 in appendix)

Source	Source 1 Angle (deg)	Source 2 Angle (deg)	Distance (m)	Reverb
Music	-50	-10	1	250 ms
Female Speech	-50	-10	1	130 ms
Male Speech	N/a	N/a	N/a	N/a

Table 1. Source configuration for each sample

After mixing these individual sources, each of the codes given from the papers were ran on the combined two channel signals. Note that although FCA is capable of handling underdetermined cases, the rest of the approaches are not, so only 2 channel, 2 speaker determined tests were ran.

For FastMVAE2, the original trained model was used. This model was trained using VCC data [18] by the paper's authors, and assumes 4 classes for the latent variables.

FastFCA and HVA were both ran on an Intel i7-5500U processor, while the Python implementation for FastMVAE2 required CUDA. I do not have a dedicated GPU, so it was ran in Google Colab, using Nvidia T4.The wallclock time for FastFCA and HVA were captured for each separation task.

Approach	Matlab Mixed Anechoic	Echoic Speech	Music
HVA	22.175027 s	54.856291 s	26.163857 s
FCA	26.834865 s	32.000792 s	19.542660 s

Table 2. Wallclock time for HVA and FCA, ran on Intel i7-5500U

The results were then manually matched between source and output to correct permutation, and metrics were calculated using the fast-bss-eval module from [14]. Details on the metrics and their interpretation can be found in A.4.

Note that the music sample was omitted for FastMVAE2, since it is specifically trained for speech separation.

Unsurprisingly, in the anechoic case FastMVAE2 has an exceptional SAR. Due to source mixing with a random weight matrix, SIR and SDR appear quite low for all methods. Visual comparison between the mixed spectrograms (2) also confirms that one source is dominant across the spectrum.

The music sample was specifically challenging to HVA, since the signals mixed were two guitars playing similar melodies. This causes HVA to struggle in isolating harmonics from one another since the sources are harmonically indistinguishable. Exacerbating the issue, the signals are heavily correlated in time as well since they are both strumming in time with each other. This issue in HVA is indicated by SIR where FCA is a little over 6 times higher than HVA. However, SDR and SAR are very similar between the two.

Approach	SAR (dB)	SDR (dB)	SIR (dB)
Male Speech (Anechoic)			
HVA	33.85	2.27	2.31
FastFCA	21.02	2.48	2.57
FastMVAE2	37.00	2.30	2.30
Music (Reverberant)			
HVA	8.71	4.55	7.30
FastFCA	9.15	6.95	15.20
FastMVAE2	N/a	N/a	N/a
Female Speech (Reverberant)			
HVA	12.68	12.02	21.77
FastFCA	11.28	8.34	13.66
FastMVAE2	12.39	10.10	14.31

Table 3. Average SAR, SDR, and SIR for different approaches across datasets.

For the reverberant speech test, HVA had by far the highest SIR. A visual inspection of the spectrograms for the mixed signals 4 indicates that the time-frequency patterns and specifically harmonics in the two speaker's voices are identifiable, meaning HVA should have a clean separation. Unsurprisingly FastMVAE2 seems to struggle with the reverberant environment relative to the anechoic environment, as it is not trained using reverberant speech, indicated by a lowered SAR. However FastMVAE2 is still able to outperform FastFCA in this case.

The metrics for each channel are provided in A.4.

7. Conclusion

In conclusion, three methods have been compared using both deep learning based approaches and statistical methods. While deep learning is still very promising, future work is needed to improve generalization performance and performance in reverberant environments. Deconvolution is still an open question that will need to be solved to realize gains in reverberant samples.

Further work must be done to improve BSS in real time. The context for BSS is largely hearing aids, which necessitate reduced computation and have limited spacing for microphones, meaning spatial resolution is not guaranteed.

More work should focus on incorporating prior information, such as beamforming weights or temporal cues to provide 'attentive listening'. Many current approaches operate unsupervised, leaving knowable information out of the solution.

References

[1] S. Ansari, A. S. Alatrany, K. A. Alnajjar, T. Khater, S. Mahmoud, D. Al-Jumeily, and A. J. Hussain. A survey of artificial intelligence approaches in blind source separation. *Neurocomputing*, 561:126895, 2023.

[2] N. Q. K. Duong, E. Vincent, and R. Gribonval. Underdetermined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1830–1840, 2010.

[3] E. A. P. Habets. *Speech Dereverberation Using Statistical Reverberation Models*, pages 57–93. Springer London, London, 2010.

[4] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430, 2000.

[5] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430, 2000.

[6] N. Ito, R. Ikeshita, H. Sawada, and T. Nakatani. A joint diagonalization based efficient approach to underdetermined blind audio source separation using the multichannel wiener filter. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1950–1965, 2021.

[7] H. Jin, W. Luo, H. Li, and L. Dai. Underdetermined blind source separation of radar signals based on genetic annealing algorithm. *The Journal of Engineering*, 2022(3):261–273, 2022.

[8] L. Li, H. Kameoka, and S. Makino. Fastmvae2: On improving and accelerating the fast variational autoencoder-based source separation algorithm for determined mixtures. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:96–110, 2023.

[9] Y. Li and D. A. Ramli. Advances in time-frequency analysis for blind source separation: Challenges, contributions, and emerging trends. *IEEE Access*, 11:137450–137474, 2023.

[10] Z. Liu, W. Liang, N. Fu, L. Qiao, and J. Zhang. Main lobe deceptive jamming suppression based on blind source separation and energy detection for monopulse radar. *IET Radar, Sonar & Navigation*, n/a(n/a).

[11] J. Malek, J. Jansky, Z. Koldovsky, T. Kounovsky, J. Cmejla, and J. Zdansky. Target speech extraction: Independent vector extraction guided by supervised speaker identification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2295–2309, 2022.

[12] A. A. Nugraha, A. Liutkus, and E. Vincent. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1652–1664, 2016.

[13] H. Sawada, R. Mukai, S. Araki, and S. Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Transactions on Speech and Audio Processing*, 12(5):530–538, 2004.

[14] R. Scheibler. Sdr — medium rare with fast computations, 2021.

[15] SISEC. Sisec 2011 database, 2011. Accessed: 11/30/2024.

[16] R. Wildeboer, F. Sammali, R. van Sloun, Y. Huang, P. Chen, M. Bruce, C. Rabotti, S. Shulepov, G. Salomon, B. Schoot, H. Wijkstra, and M. Mischi. Blind source separation for clutter and noise suppression in ultrasound imaging: review for different applications. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 67(8):1497–1512, Aug. 2020.

[17] K. Yatabe and D. Kitamura. Determined bss based on time-frequency masking and its application to harmonic vector analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1609–1625, 2021.

[18] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z.-H. Ling, and T. Toda. Voice Conversion Challenge 2020 — Intra-lingual semi-parallel and cross-lingual voice conversion —. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pages 80–98, 2020.

[19] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, July 2004.

A. Appendix

A.1. Definitions

Definition A.1. Let $\mathbf{X}$ be a random variable with PDF f. The differential entropy $h(\mathbf{X})$, or $h(f)$ is:

$$h(\mathbf{X}) = \mathbb{E}[-\log(f(\mathbf{X}))]$$

Theorem A.2. *Differential entropy for a given mean and variance is maximized by a Gaussian distribution.*

Proof. Let $g(x)$ be a Gaussian PDF with mean μ and variance σ^2, and $f(x)$ is an arbitrary PDF with the same mean (μ) and variance (σ^2).

Considering the Kullback-Leibler divergence between the two:

$$0 \leq \mathbf{D}_{KL}(f||g) = \int_{-\infty}^{\infty} f(x)\log(\frac{f(x)}{g(x)})dx = -h(f) - \int_{-\infty}^{\infty} f(x)\log(g(x))dx$$

Where $h(f)$ is the differential entropy of the PDF f, defined in A.1. Next:

$$\int_{-\infty}^{\infty} f(x)\log(g(x))\,dx = \int_{-\infty}^{\infty} f(x)\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)dx$$

$$= \int_{-\infty}^{\infty} f(x)\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)dx + \log(e)\int_{-\infty}^{\infty} f(x)\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx$$

$$= -\frac{1}{2}\log(2\pi\sigma^2) - \log(e)\frac{\sigma^2}{2\sigma^2}$$

$$= -\frac{1}{2}(\log(2\pi\sigma^2) + \log(e))$$

$$= -\frac{1}{2}\log(2\pi e\sigma^2)$$

$$= -h(g)$$

Note that $0 \leq \mathbf{D}_{KL}$, so substituting back:

$$-h(f) + h(g) \geq 0$$

$$h(g) \geq h(f)$$

With equality where

$$f(x) = g(x)$$

□

Definition A.3. For a continuous time signal, $x(t)$, the power spectral density (PSD) $S_x(f)$ is defined as the Fourier transform of the autocorrelation function $R_x(\tau)$ of the signal, where the autocorrelation $R_x(\tau)$ is given as:

$$R_x(\tau) = \mathbb{E}[x(t)x(t+\tau)]$$

Where $\mathbb{E}$ is the expectation operator, and τ is a time lag. The power spectral density is a quantification for how the total power in a signal is distributed across frequencies. Some properties of power spectral density include:

- **Non-negativity**: $S_x(f) \geq 0$ for all frequencies f
- **Integration**: The total power in a signal, P, is given as:

$$P = \int_{-\infty}^{\infty} S_x(f)df$$

Definition A.4. Let the received signal $x(t)$ be modeled as:

$$x(t) = s(t) + n(t)$$

Where $s(t)$ is the desired signal, and $n(t)$ is a noise term. Let the Power Spectral Density (PSD) of $s(t)$ be denoted as $S_{ss}(f)$, the PSD of $n(t)$ be $S_{nn}(f)$. Assuming that the signal and nosie are uncorrelated, the wiener Filter is then defined as:

$$H(f) = \frac{S_{ss}(f)}{S_{ss}(f) + S_{nn}(f)}$$

Where the output of the Wiener Filter is defined as:

$$Y(f) = H(f) \cdot X(f)$$

Where X(f) is the Fourier transform of the observed signal x(t). The Wiener filter is typically updated using the mean-squared error between $y(t)$, the inverse Fourier transform of $Y(f)$, and the signal $s(t)$:

$$MSE = \mathbb{E}[||s(t) - y(t)||^2]$$

A.2. Method Comparison

Feature	**FastFCA**	**HVA**	**FastMVAE2**
BSS Scenario	Underdetermined + Determined	Determined	Determined
Core Technique	Joint Diagonalization+ Time-Frequency Masking	Time Frequency Masking	Deep Generative Model
Exploited Property	Spatial Covariance	Harmonic Structure	Spectrogram Patterns

Table 4. Comparison of features across FastFCA, HVA, and FastMVAE2.

A.3. Spectrogram Plots of Source Signals

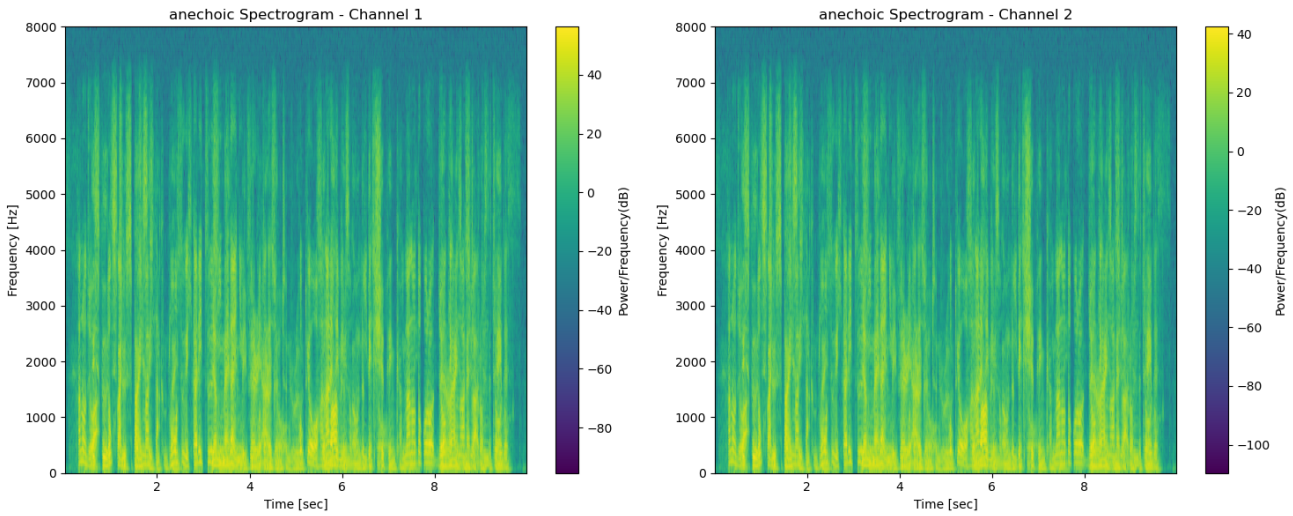


Figure 2. Spectrogram of Mixed Anechoic Sources

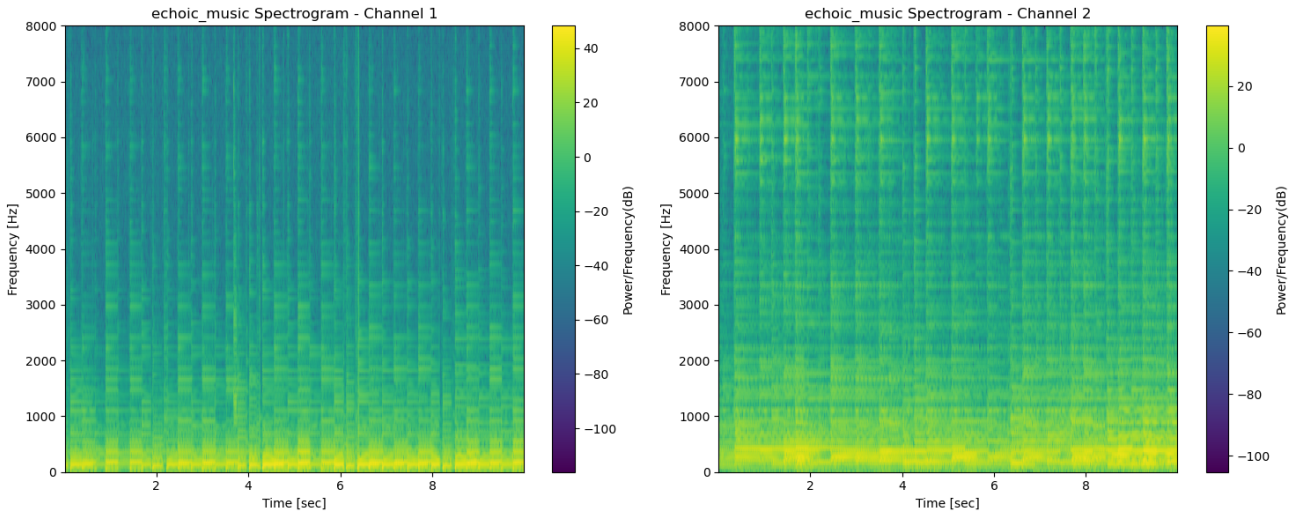


Figure 3. Spectrogram of Mixed Reverberant Music

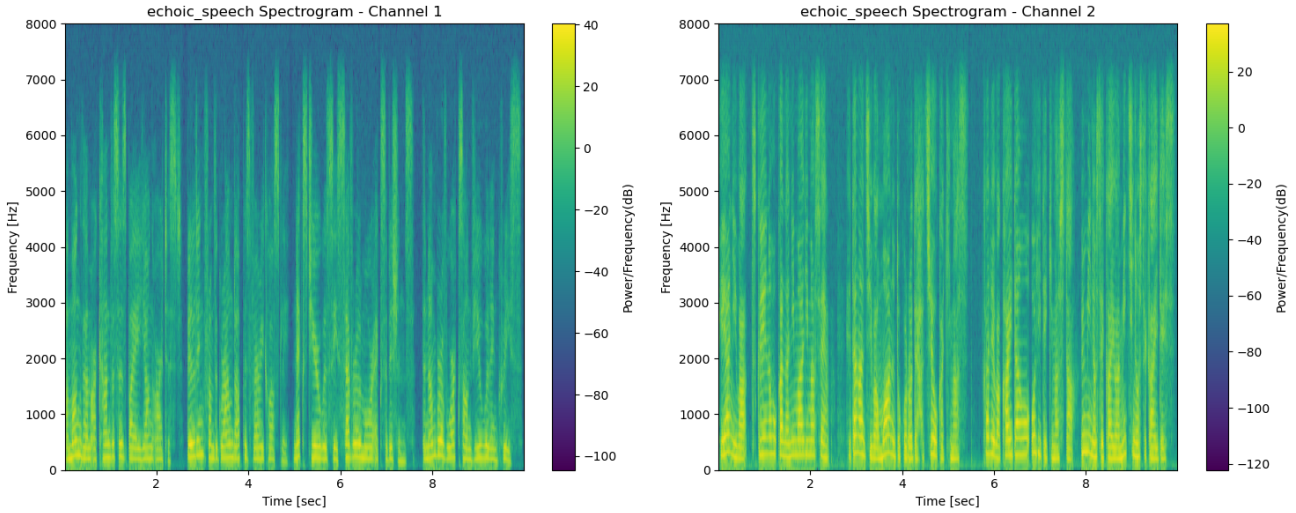


Figure 4. Spectrogram of Mixed Reverberant Speech

A.4. Blind Source Separation Metrics and Their Interpretation

BSS metrics are often used to describe separation performance. In general, three metrics are typically used: Signal to Artifact Ratio (SAR), Signal to Distortion Ratio (SDR), and Signal to Interference Ratio (SIR). These three metrics, often reported in decibels (dB), give a quantification for how well sources have been isolated and recovered.

Definition A.5. SAR is defined as:

$$\mathrm{SAR} = 10\log_{10}\frac{\|s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}}\|^2}{\|e_{\text{artif}}\|^2},$$

Where e indicates an error and s indicates the true source signal. SAR is interpreted as how well the algorithm has recovered the signal. In other words, it indicates the relative power level between 'organic' signal and 'artificial' signal left from the separation process.

Definition A.6. SDR is defined as:

$$\text{SDR} = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}\|^2},$$

Where e indicates an error signal and s indicates the true source signal. SDR is an indication of the overall BSS performance, as the goal is to maximize signal and suppress errors.

Definition A.7. SIR is defined as:

$$\text{SIR} = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2},$$

Where s is the true source signal, and e is an error signal. SIR is an indication of how corrupted the recovered signal is by other signals. It is a measure of how much of the signal is affected by other signals.

Dataset	Approach	SAR S1 (dB)	SAR S2 (dB)	SDR S1 (dB)	SDR S2 (dB)	SIR S1 (dB)	SIR S2 (dB)
Male Speech (Anechoic)	HVA	24.87	36.58	2.16	2.39	2.19	2.42
	FastFCA	21.43	20.57	2.37	2.58	2.45	2.68
	FastMVAE2	37.39	36.58	2.17	2.42	2.18	2.42
Music (Reverberant)	HVA	7.35	9.74	1.99	6.15	4.22	9.09
	FastFCA	8.65	9.60	3.11	8.95	5.08	17.99
	FastMVAE2	7.04	9.30	1.14	7.36	3.22	12.26
Female Speech (Reverberant)	HVA	10.48	14.13	9.22	13.71	15.59	24.22
	FastFCA	11.79	10.70	6.81	9.47	8.76	15.90
	FastMVAE2	13.85	10.19	11.40	8.24	15.28	13.06

Table 5. Detailed breakdown of SAR, SDR, and SIR for each source and approach across datasets. Averages are omitted for clarity.

0.125	0.768
0.05	0.864

Table 6. Anechoic mixing matrix given by rand([2,2]) in MATLAB

A.5. Code Used

Code was pulled from the respective repositories from the papers cited ([17],[8],[6],[14]). These can be found here:

- **HVA**:

  `https://codeocean.com/capsule/2232702/tree`

- **FastFCA**:

  `https://github.com/nttcslab-sp/unifiedUdetDetBSS/tree/master`

- **FastMVAE2**:

  `https://github.com/lili-0805/mvae-ss/tree/main`

- **BSS Metrics**:

  `https://github.com/fakufaku/fast_bss_eval/tree/main`

A.6. Knowledge Distillation

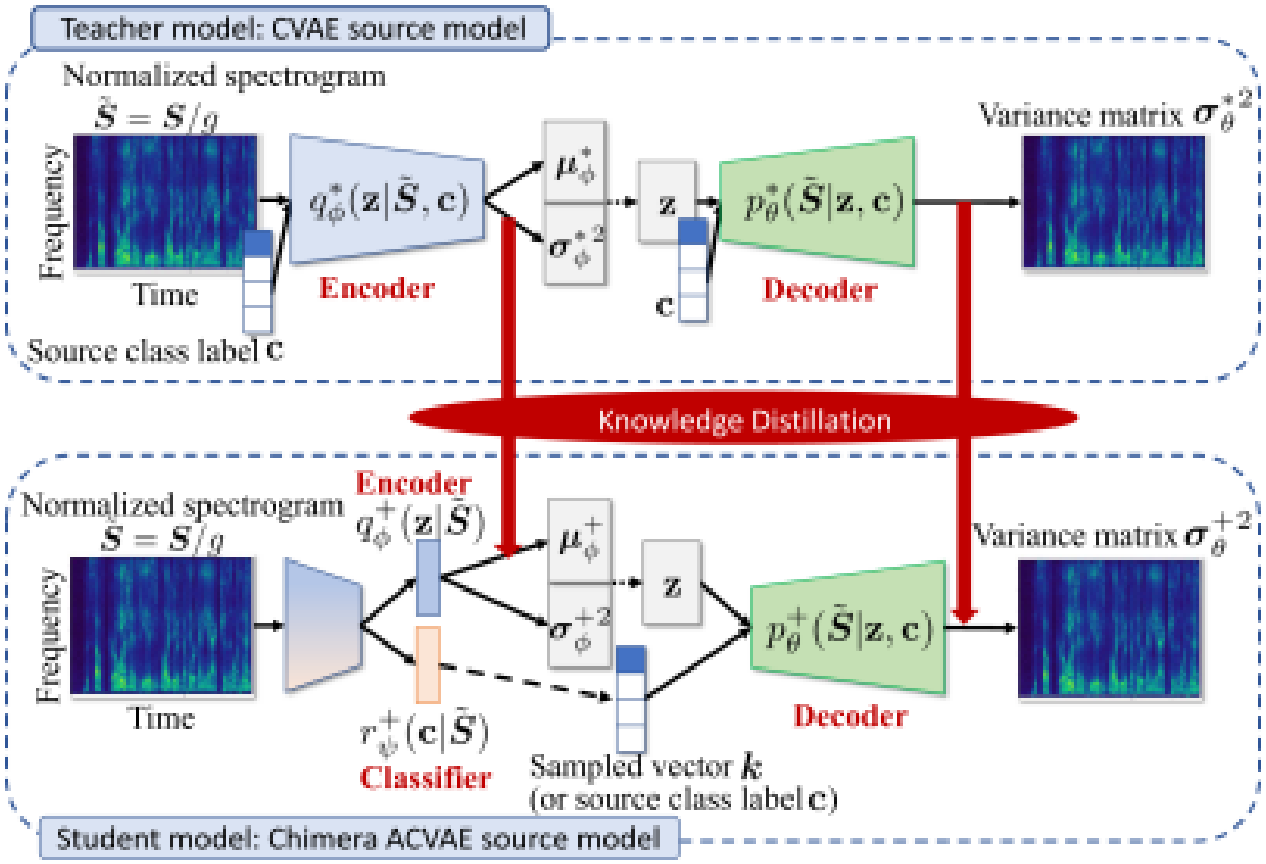


Figure 5. Training Knowledge Distillation Process for ChimeraACVAE