International Journal on Emerging Technologies 11(3): 927-935(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Non-Negative Matrix Factorization based Blind Source Separation and Source Enhancement using Generalised Cross Correlation

R. Pradeep[1], R. Kanimozhi[2], C. Prajitha[3] and S. Rinesh[4]

[1]Assistant Professor, Department of ECE, Sri Eshwar College of Engineering, CoimbatoreTamil Nadu, INDIA.
[2]Junior Research Fellow, Centre for Interdisciplinary Research, Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, INDIA.
[3]Junior Research Fellow, Centre for Interdisciplinary Research, Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, INDIA.
[4]Assistant Professor, Department of CSE, Jigjiga University, Ethiopia

(Corresponding author: R. Pradeep)
(Received 30 March 2020, Revised 01 June 2020, Accepted 03 June 2020)
(Published by Research Trend, Website: www.researchtrend.net

ABSTRACT: The use of speech recognition technology for different applications is required widely, such as protection, translation of the word to text, etc. The initial signal that degrades the output of the ASR network incorporates certain unnecessary signals from different sources during speech signal acquisition. This problem leads to the development of a robust algorithm for audio demixing or separation of the source. This is knelt like sounds or mixing of sources, due to multi-user recording and echo effect. It is quite difficult to remove such mixed sounds in many domains like crime, evidence, etc. This remains as a major challenge in separating such sounds. To overcome this problem, a new methodology for Blind Source Separation is proposed using a Non-negative Matrix Factorisation using Generalised Cross-Correlation (BSS: GCC-NMF) to address this issue in this article. It is proved that Generalised Cross-Correlation – Non -negative Matrix is sound source isolation, excellent including both blind source and several knowledgeable source approaches involving advanced knowledge. The proposed GCC is used to improve the extraction of specific audio signals and also it can reduce the time delay in the speech separation process. Similarly, NMF is applied to split the entire sound signal into factorized matrix without low pitch signals which is represented as negative elements. So, it has removed the negative elements in the respective matrix. Experimental studies show the strength of the proposed model by analyzing different scenarios of implementation. Various performance evaluation has been done for both divergence and distance-based conditions and the proposed method achieves higher performance rate.

Keywords: Speech Recognition, Speech Signal Acquisition, Multi-user Recording, Blind Source, Generalised Cross-Correlation, non-negative Matrix Factorisation

Abbreviations: BSS, blind source separation; GCC, generalised cross correlation; NMF, non-negative matrix factorisation; PCA, principle component analysis; ICA, independent component analysis; CPP, cocktail party issue; SAR, sources to artifacts ratio; SDR, sources to distortion ratio; SIR, Sources to interference ratio.

I. INTRODUCTION

Hearing aids, ASR, and many other communications systems work rather well when one source is almost without an echo [1]. Still, in situations where acoustic sources are simultaneously generated, or vibrations are high, their performance deteriorates [2]. Therefore, it is very convenient, using a front-end model as used in the cocktail issue, to localize and isolate the source signals [3]. It is interesting to note that humans and owls can detect and localize sound stimuli to a certain degree [4]. The exact location of fruit and predators is calculated by bats using the concept of echolocation [5]. In the face of other speakers or disturbances, individuals will concentrate on a single voice, demonstrating an incredible capacity to focus on one source and cancel the other speakers [6,7]. Nevertheless, in comparison with human performance, machinery performance is rather low [8].

Active audio sources include mixtures of different speakers and music concerts with the signals, which are mixtures of musical instruments and sound recordings [9]. Many audio signals are concurrently active blends of several audio outlets. The feature of human hearing is the capacity to differentiate between discrete sound signals and complicated sound mixtures [10]. This is related to the well-known problem of cocktail parties. When there are many conversations and background noises, the cocktail party is characterized as the focus of attention on one speaker [11].

Late, array analysis approaches have traditionally been used in the Blind Source Separation (BSS) processes [12-14]. The separation was achieved by assessing the spatial location of the sources and using a fitting spatial filter to raise the target source and eliminate the other sources from the mixture channels [15,16]. Such techniques involved the capturing of sources in a complicated setup with a large number of microphones [17]. This restricted its implementation to some live broadcasting circumstances in the conference auditorium, for example, in which the microphone configuration was changed at will. The computational signal processing and computer sciences jointly suggested BSS approaches appropriate for reaction mixtures and more conventional documentation [18-20]. An early approach had an algorithm for the neural network, which could extract separate sources using a kind that was not linear. Many scholars have researched the use of the Independent Component Analysis (ICA) algorithms in a computational context, mainly based on higher-order statistics and theory of results. These new algorithms were confined to immediate mixtures, which

contain as many sources as mixing channels [21]. Through expanding ICA into uncertainty mixture, the first move to cope with possible audio combinations was taken. Some students have, in the meantime, researched perceptive sound processing concepts arising from listening tests and created Computer Auditory Scene Analysis algorithms [22]. The basic properties of the audio inputs, including harmonicity or spectral envelope, were used in these algorithms to decipher single channel mixtures into distinct sensory streams that could be resynthesized separately [23]. The approaches used to evaluate the early computer auditory environment used a series of data-driven processing steps, whereas the latter proposed that a prediction-driven methodology would be used as a blackboard model. Over the last decade, systems such as ICA and ICA dependent CASA, and modern BSS methods have been introduced [24].

The main objective of the study is to retrieve root signals for cocktail parties and multi-speaker environments from the obtained mixtures. If more than one speaker talks at the same moment, the microphones capture the sounds from all speaker [25]. Consequently, the reported signals known as mixtures include origins of different time delays of different proportions. Every reported combination consists of the sources of different parts, provided the microphones located at various locations. You may want to select a specific destination or collect all the numerous speech sources.

Specific BSS implementations provide real-time isolation of speeches for simultaneous and electronic music processing support for video sampling [26]. More specifically, other related systems are programmed to change the mixing signal by varying mixing of the sources or eliminating unacceptable sources. These remix applications include Speech enhanced hearing aids and mobile phones, karaoke voice cancelation, multi-channel stereo CD rendering, raw music post-production, or the restoration of corrupted audio data. Single-source indexing, encoding, and coding techniques can also be used to facilitate other applications, such as multiple-source tracking and robotics place, improved automated audio-document indexing, multi-speech recognition in cocktail parties or object-specific coding [27].

For digital signal processing, one of the essential criteria is to retrieve the initial signals from the signal mix. The method by which the initial signals are measured and separated from the mixture is called the root isolation. A common source isolation situation is the cocktail party, where we have many speakers, background music, and vibration, and only the combination is understood [28]. Still, the primary source of interest is calculated, or even extracted from the microphone sound-mixtures. Blind-source segregation or blind-signal separation (BSS) is the method for distinguishing the signals without prior knowledge of the signal mixing systems [29].

Principal component analysis, single values decomposition, separately evaluated object analysis, NMF, stationary subspace analysis, and specific spatial structures are fundamental methods of source separation. These are the main methods of source separation [30]. Such approaches are used for the study of mixed data for optimal separation performance, depending on the type of source separation issue and the form of mixing device involved [31]. BSS is a space technology which has been established by audio processing but also can be extended to image processing, biomedical details, telecommunications, etc. [33]. The various related study and techniques are viewed in this section.

L. Deng *et al.* elaborated that Speech, music or noises are usually known as the audio source [34]. That type has its features that can be used for a particular process. Word sounds can be perceived as a series of distinct phonemes. Due to the co-articulation of succession phonemes, the signal matching each phoneme exhibits time characteristics .These signals may include a periodic part of harmonic sinusoidal sections induced by regular vocal cord movements, the broadband sound of passage of air through mouth, or the intermittent portion created by the pressure released suddenly behind lips or teeth. A lot of phonemes comprise regular and disruptive components in superposition. The harmonicity property means that the sinusoidal component frequencies are many times the same rate considered the fundamental frequency. The critical frequencies of the standard phonemes differ due to intonation but usually remain within 42 Hz, concentrating on males about 142 Hz and female speakers around 200 Hz. The spectral envelope is determined by the structure of the vocal tract, which relies on phoneme, phonetic meaning and speaker features, and is the smooth amplitude of the signal as a frequency feature. Successive phonemes are composed of words and phrases which rely on language and the lexical and semantic laws.

E. colin termed that Colin Cherry coined the cocktail party issue (CPP) [35]. It is a psychoacoustic condition that relates to the extraordinary human capacity to listen, track, and selectively perceive one sensory stimulus in a noisy environment. Whether overlapping speech sounds or several stimuli which are often supposed to be independent of each other, contribute to auditory disturbances. The numerous details concerning CPP were investigated by Simon Haykine [36]. The CPP is defined as the issue of such a mixed recording and distinguishing the individual speakers. The generalization of this issue is known as the blind source distinction between the various independent components of a signal, without using any particular knowledge of the component signals.

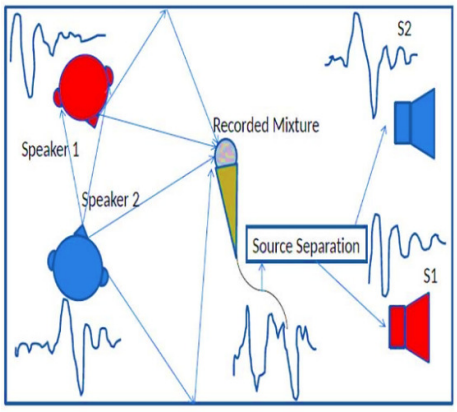

Fig. 1. A more straightforward cocktail question situation of two speakers and a combination recording.

Herault *et al.* introduced that Without previous knowledge regarding mixing processes or source signals, the question of isolation from their observable composition of unobserved sources is regarded as the blind source segregation [37]. Remember, for example, as in Fig. 1, the situation of two people talking with two microphones positioned in two different positions

recording the mixed signals. The Blind Speech Separation algorithm here is designed to separate speech signals from mixed signals obtained from microphone output without prior knowledge of the source, microphone position, or mixing processes, i.e., the blind distance between source and mixture. In this specific example, the details are combined acoustically, and the question is commonly known as the problem for a cocktail party. Then other algorithms were designed to separate the signals from their essential simultaneous mixes or their complicated non-linear, convolutive time-variant variations.

Different separating methods are given depending on the underfunded or overfunded program and also on the number of microphones or speakers that we have. Recently unattended machine learning algorithms of single-channel source isolation have been used effectively. Typically based on the single linear model, the separation takes place by finding a decomposition in which the sources are statistically independent or non-redundant instead of using previous knowledge of the source signal features. Algorithms based on independent analysis of components (ICA), non-negative matrix factoring, and sparse coding were proposed, as shown in Fig. 2.

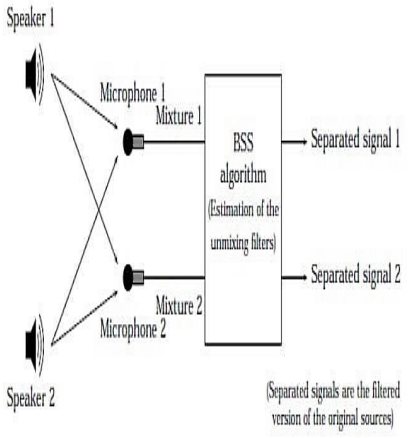

Fig. 2. Illustration of the blind source separation problem.

It is a statistical method for the decay of additive subcomponents of a multivariate signal. ICA may be used to capture signals from the combination of other signals and vibration in intense listening environments with scattered microphones. Sub-components should be non-Gaussian and statistically stable, all requirements of ICA decomposition. Bell proposed the concept of the ICA algorithm [38].

He concentrated on approaches to the instantaneous mixing issue, which implies that noise-free and linear combinations of component signals were chosen for the mixed signal.

Paatero and Tapper presented NMF and named it a constructive factoring matrix, and subsequently, Lee and Seung researched the propensities of the algorithm. The NMF is a non-negative limit decomposition strategy [39].

Virtanen et al. suggested an NMF-based unmonitored algorithm for sound source isolation that blends NMF with time duration and slimming goals [40] . A higher separation efficiency than the current algorithms has been shown by the algorithm suggested. Schmidt proposed using sparse, non-negative matrix factorizations to make the sparse coding separation computationally attractive. As a first step, the fragmented description of the signals is evaluated in full dictionaries for different speakers.

The output signals are divided by the mixture of the dictionaries related to the source and the coarse deterioration measured. It explores the value of the degree of sparsity and the number of dictionary components. The simple unattended SN is then contrasted to an algorithm-controlled program in which the details are separated into phoneme-level sub-problems, which result in significant code savings.

For speech separation in verbal communication it seems to be more detrimental in implantation. Likewise, in some research, several sensors were used. Whereas in that condition the sensor numbers were too less than the signals.

Another method called Equal Input and Equal Output for separation also founds a drawback that it could not be applied in common pedagogical situation. In all the research, some environmental sounds disturb the output.

Based on the survey, several issues related to speech processing has been studied, To overcome such problems. The new method BSS-GCC-NMF is discussed as follows:

“The Proposed technique can overcome the practical issues by factorize the speech signal and by reduce the time delay of the process.”

II. METHODOLOGY

The system consists mainly of six sections: the continuum of the speech sound, decomposing the signal, finding the source, extracting the voice, reconstruct the voice, and improving the quality of the expression at last. The NMF signal is used to decompose. The combination speech signal distribution is calculated by adding Short Time Fourier Transform before the decomposition is done. For source localization, the Generalised Cross-Correlation Step Transform algorithm is used to predict the arrival time interval. The origins are then isolated by the use of the coefficient masking procedure. Ultimately, it is reconstructed via reverse NMF and reverse FT sources. Spectral subtraction and adaptive filtering are used to enhance the quality of separate sources. The frame is used for speech combination isolation of certain stereotypes. All the voice signals, both the mixture and the source signals, are sampled at 20 kHz in the data set. In this case, IT is used under specified Live Speech Mixtures. In this paper, the block diagram showing the technique of speech separation is given in Fig. 3.

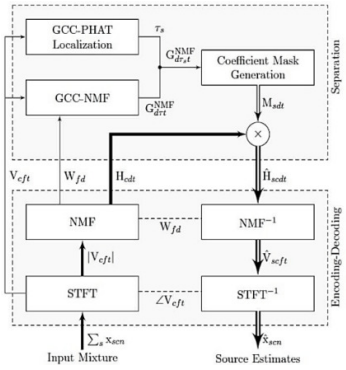

Fig. 3. Block Diagram of the methodology adopted for source separation.

A. Spectrogram estimation

For calculating the range of the provided audio signal, Short Time Fourier Transform is used to measure the long-term audio signal in shorter sections of equivalent distances, and the Fourier will then be extracted from each portion of the audio signal.

– Set window for review
– Defines the quantity of window overlap
– Enable a windowing option
– Generate window segments (signal multiply by a window)

In this paper, the spectrogram is measured using the Hanning window function. The audio signal is split into 1024 different segments, with 128 samples alternating two consecutive layers. The left and right channels are measured independently because the audio is a stereotype. Fig. 4 displays the audio combination signal.

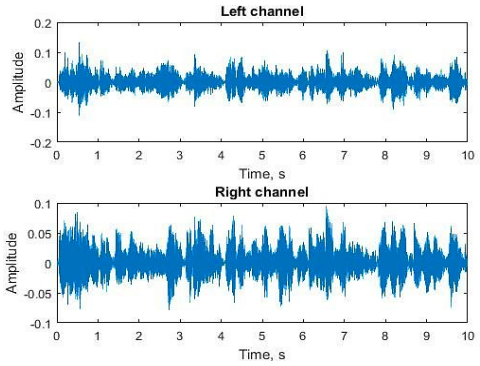

Fig. 4. Mixture audio signal dev1 female3 live rec 130ms 1m mix.wav.

Apply the Fourier Transform on every window

Hanning window represents the function as ibn equation (1),

$$u[x,T] = 0.56 - 0.45\cos(\frac{2\pi(x-T)}{X_u - 1} \quad (1)$$

The Short-Time Fourier Transform is given as in equation (2),

$$W(x,u_n) = \sum_{n=-\infty}^{\infty}(u(n)e^{-ju0n})u(x-n) \quad (2)$$

Algorithm:

1. *L* and Bare initialized with the constraint L *B* >0 randomly.
2. Update Lusing the updatingrule

$$L \leftarrow L \odot \frac{(Z^2 \odot U)B^T}{Z^{-1}B^T} \quad (3)$$

3. Update *B* using updatingrule

$$B \leftarrow B \odot \frac{L^T(Z^2 \odot U)B^T}{L^T Z^{-1} + \propto} \quad (4)$$

4. Optimal *L and B* are obtained when the *β* - divergence $D_\beta(U|Z)$ is reduced.

$D_\beta(U|Z)$ is given by equation (5),

$$D_\beta(U|Z) = \begin{cases} \frac{U}{Z} - \log\left(\frac{U}{Z}\right) - 1 & if \ \beta = 0 \\ U(\log(U) - \log(Z)) + (Z - U) & if \ \beta = 1 \\ \frac{1}{\beta - 1}\left(U - Z + \beta Z - 1(U - Z)\right) & \text{otherwise} \end{cases} \quad (5)$$

— *β* = 0: Itakura Saitodivergence
— *β* = 1: KullbackLeiblerdivergence
— *β* = 2: Euclidiandistance

B. Signal decomposition

Optimal properties are extracted using non-negative matrix factorization from the amplification spectrogram obtained following the application of Short Times Fourier Transform. It is used to break down a large, not negative matrix into less-dimensional, non-negative factor matrices. Two parts of Spectrometer U have spectrogram left and right, and both have measurements. The amplitude Specification U has two items. Two element matrices L and B are obtained following the application of Non-negative Matrix Factorization on U. L is defined as the vector for a dictionary and B as a matrix coefficient. L and B are such that Z approximates U of their output.

C. Source localization

Time Arrival differences are an essential localization parameter. And a signal structure of the time-delay so period t is calculated from the angular spectrogram. Various angular spectrogram measurement techniques are usable. Here is the method Generalized Cross Correlation-Phase Transform.

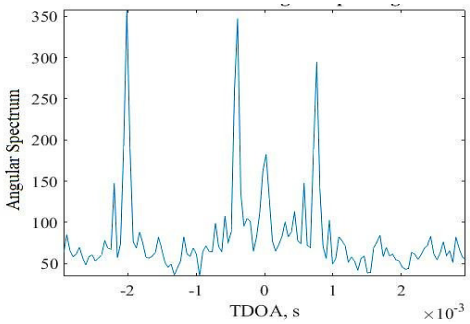

Fig. 5. Mean angular spectrogram.

The angular spectrum is represented as in equation (6)

$$Dxt = Y\psi ft * \frac{e^{j2\upsilon ft}}{f} \quad (6)$$

The generic cross-correlation and not negative matrix factorization are used to distinguish the source. Since the non-negative atoms of the matrix factorization dictionary itself are non-negative frequency functions, this aspect may be used to establish a set of frequency-weighting features of the generalized cross-correlation of bits of the normalized dictionaries.

Source-specific binary represented as in equation (7)

$$W_{sdT} = \begin{cases} 1 & if \ s = 1 \\ 0 & otherwise \end{cases} \quad (7)$$

D. Source separation

The combined GCC and NMF is used for proper source separation for source separation a combination of GCC and NMF is used. Since NMF dictionary atoms are themselves non-negative functions of frequency this fact can be used to define a set of GCC frequency-weighting functions φ_{dft}^{NMF} from the normalized dictionary atoms as in equation (8),

$$\varphi_{dft}^{NMF} = \frac{1}{|V_{ift}||V_{rft}|}\frac{W}{\sum_f W} \quad (8)$$

Atom-specific angular spectrograms is then defined by equation (9),

$$W_{dft}^{NMF} = \sum \varphi_{dft}^{NMF} V_{ift}\, e^{j2\pi f} \quad (9)$$

The atom-specific angular spectrograms defined above can be used to associate each atom, at each point in time, to a single source based on its spatial origin. Similarly, rom the TDOA values source scan be localized and to separate the source from the mixture signal coefficient masking can be performed. For each source a binary mask M_{dt} is assigned, whose values will be either 1 or 0, according to the angular spectrogram and the dictionary matrix *W*. At each point of time it is checked that the atom *d* in *W* belongs to which source and accordingly the value is set as 1 for that particular source. The coefficient mask of other sources will be set as 0 at that point of time. As a result, each atom is attributed to a single source at each point of time.

E. Source reconstruction

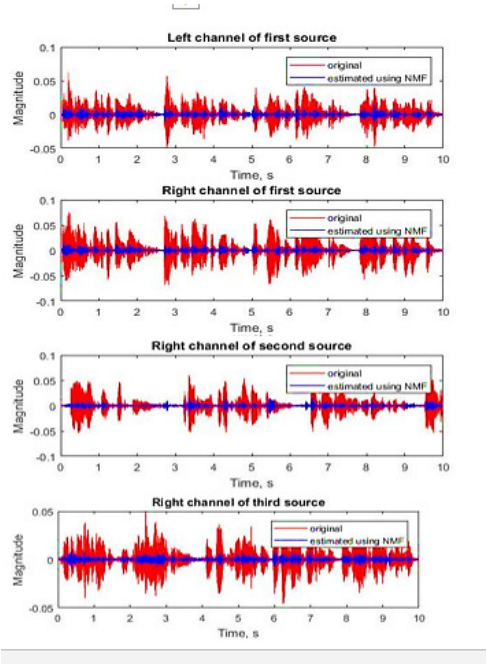

Fig. 6. Original and reconstructed sources 1, 2 and 3.

A stereo source approximation spectrogram may then be calculated by multiplying the mask by the non-negative factoring matrix dimension before reconstruction. The Inverse Short Time Fourier Transform is used to transform the individual source spectrum into the time domain and it is given as in equation (10).

$$X^{scn} = STFT - 1\ (U^{scft}) \quad (10)$$

F. Speech enhancement

The segregated speech streams may also involve intervention from other sources following the implementation of non-negative matrix factorization. By removing the effects of different sources from each source, the quality of the separate speech signal can be improved. Improvement is made for this voice. Below are two forms of speech enhancement.

Adaptive filtering. In this case, the order 11 LMS filter is used to eliminate adaptive noise. The stage is adjusted to 0.01, and the tape is set to 40. After enhancement with adaptive filtering, the initial sound signals and different speakers are shown in the chart.

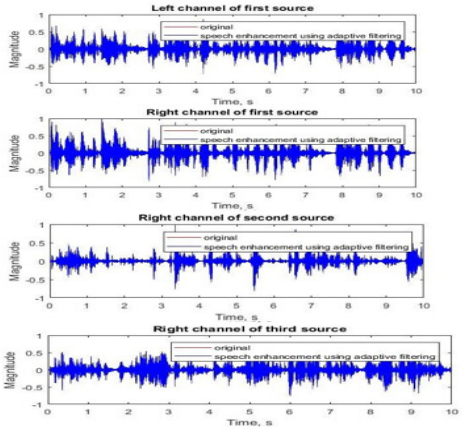

Fig. 7. Original and enhanced sources 1, 2 and 3 using adaptive filter.

Spectral subtraction. With spectral subtraction, the noise spectrum is evaluated by averaging the speech pause frames and is deduced from the noise signal spectrum. Every separate source signal is a sum of the desired source signal and interferences of others. The resultant signal is then subject to half a wave correction rule in order to negate the negative values.

Non-negative matrix factorization (NMF) NMF is a category of matrix decomposition algorithms with a non-negative limit, which is to be factorized and which should not cancel the matrix factors. NMF is a universal data decomposition technique. Through preprocessing signaling and labeling, NMF has identified a wide range of applications.

KL Divergence is to be minimized between the input data matrix U and the input data matrix estimate Z as in equation (11). It is done iteratively by updating L and B until the costs converge. Simultaneously L and B should be revised. We need to change the corresponding Column B after one row of L is modified.

$$Z = LB \quad (11)$$

The cost function is given by equation (12),

$$\min D\ (u|LB), LB\ \geq 0 \quad (12)$$

For any two matrices, A and B divergence is given by equation (13),

$$D(I|J) = \sum_{i,j}(I_{ij} \log\left(\frac{I_{ij}}{J_{ij}}\right) - I_{ij} + J_{ij} \quad (13)$$

The multiplicative updating rules to obtain optimal *W* and *H* are given by equations (14 and 15),

$$W_{i,a}^{k+1} \leftarrow W_{i,a}^{k} \frac{\frac{\sum_m H_{a,m} V_{i,m}}{WH_{i,m}}}{\sum_n H_{a,n}} \quad (14)$$

$$H_{a,m}^{k+1} \leftarrow H_{a,m}^{k} \frac{\frac{\sum_i W_{i,a} V_{i,m}}{WH_{i,m}}}{\sum_k H_{k,a}} \quad (15)$$

Enhancement of Speech aims at improving the quality of Speech through different algorithms. Performance can be clarification and interpretation, enjoyment or continuity with some other form. Removing background noise, echo reduction, and selectively adding other levels into the voice output are the key ways to improve expression. Enhancing expression is a typical problem for two reasons:

— The characteristics of Speech will change dramatically during and within implementations if the speaking signal is distorted by noise, and it depends on the quality and features of the noise signal. But seeking algorithms in various functional conditions was very challenging. It is difficult.

— The architecture of the algorithm slows implementation, and the reliability of the algorithm for each program can, therefore, be different.

The essential y(n) input is the input signal, which includes an initial d(n) signal, plus a d(n) noise. The vibration is, indeed, contrary to the original message. In the Equation (16), the central y(n) input is specified.

$$y(n) = s(n) + d(n) \quad (16)$$

The error signal is indicated with Equation, and the approximate r(n) sound noise is shown with Equation (17 and 18), which is the inclusion of the necessary s(n) voice signals and other communication interferences d(n).

$$e(n) = y(n) - r(n) \quad (17)$$

$$y(n) = s(n) + d(n) \quad (18)$$

The spectral subtraction algorithm is used only for eliminating white noise. For eliminating all white noise and light noise Multi-band, spectral subtraction is used

Let Consider a Pure speech signal with added independent additive noise as in equations (19 and 20)

$$y[n] = s[n] + d[n] \quad (19)$$

$$Y\ (\omega, k)\ = S(\omega, k) + D(\omega, k) \quad (20)$$

The assumption is that the speech signal is divided into frames for the simplification and the Frame number is denoted by k, Since ir is already assumed that the input speech signal is not correlated with the noise at the background The power spectrum of y[n] has not contain any cross-terms, then it is as in equation (21)

$$|Y(\omega)|^2 = |S(\omega)|^2 + |D(\omega)|^2 \quad (21)$$

By subtract the received signal's noise estimate, the speech signal is given as in equation (22)

$$|\hat{S}(\omega)|^2 = |Y(\omega)|^2 - |\hat{D}(\omega)|^2 \quad (22)$$

In comparison, the crucial factor is the proper iteration number that impacting the efficiency of the speech amplification method. The higher iteration number, thus, leads to a more significant improvement in expression with less residual noise.

III. RESULTS AND DISCUSSION

Experimental setup. The algorithm is performed on the SiSEC 2016 dev1 underdetermined Live Speech Mixtures Dataset. The dataset contains stereo WAV audio files that can be imported into MATLAB using the audio read command. It includes both instantaneousmixturesandliverecordings.Instantaneous mixturesaregenerated through scaling static sources using positive gains. Both male and female speech sources are played through loudspeakers in a meeting room, recorded one at a time by a pair of omni directional microphones, and subsequently added together to create live recordings. The room dimensions for live recordings are (4.45x3.55 x2.5m). The whole process is executed in Windows 10 PC with a 64-bit processor and 4 GB internal RAM. The program is implemented using MATLAB R2017a. Iteration count of NMF is taken as 100 to get optimal factor matrices *W* and *H*. Also, the factorization rank is chosen as 128. An adaptive LMS filter with order 11 and step size 0.01 is taken for speech enhancement. Performance metrics are measured in decibels.

A. Sources to Artifacts Ratio(SAR)

The numerical artifacts in each source estimate is given by equation (23)

$$SAR = \frac{10log_{10}||S_t+C_i+C_n||^2|}{C_a} \quad (23)$$

B. Sources to Distortion Ratio(SDR)

Since the separation phenomena is global measure, the whole performance of source separation is given as in equation (24)

$$SDR = \frac{10log_{10}||S_t||^2|}{||C_a+C_i+C_n||^2} \quad (24)$$

C. Sources to Interference Ratio(SIR)

The interference level in each source due to other external source is given by equation (25)

$$SIR = \frac{10log_{10}||S_t||^2|}{||C_i||^2} \quad (25)$$

Three performance measurements considered in this project, SAR, SDR, and SIR, are listed in Table 1 to Table 5. Comparison of the BSS performance measurements when the separation is performed using Basic NMF with KL Divergence based cost function and Euclidean Distance based cost function is given in Table 1 to Table. 3represents the BSS performance evaluation when separation is performed using different types of NMF like Basic NMF, Sparse NMF, and Orthogonal NMF with Euclidean Distance based Cost function and KL Divergence based Cost function respectively. BSS performance evaluation after speech enhancement using adaptive noise cancellation and spectral subtraction is given in Table 4 and 5 compares BSS evaluation with and without speech enhancement, and Table 6 compares BSS performance of various prominent methods.

Fig. 8 to 14 represents the comparison between sources, estimated sources using NMF, and enhanced sources using adaptive filters and spectral subtraction. Fig. 14 shows the performance of BSS with and without speech enhancement.

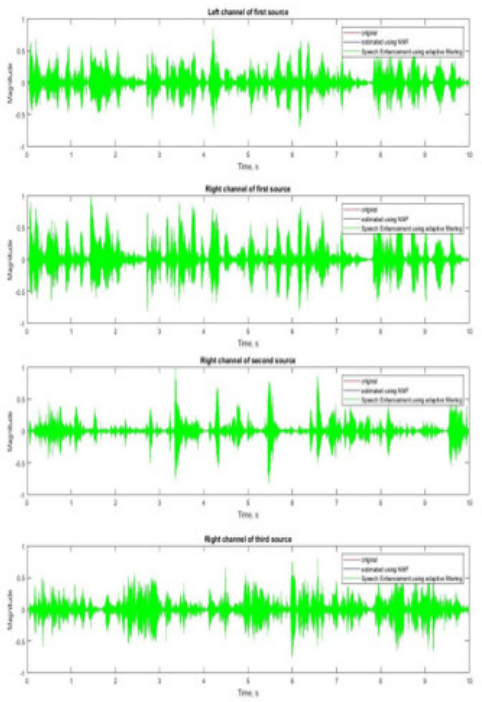

Fig. 8. Adaptive filtering output.

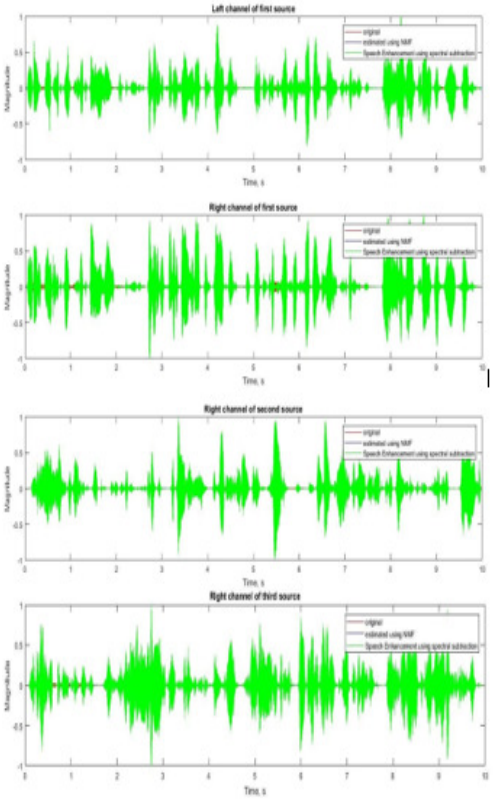

Fig. 9. Spectral subtraction output.

Table 1: Comparison of BSS performance using Basic NMF with KL Divergence and Euclidean distance-based Cost functions.

NMF Algorithm	SAR	SDR	SIR
Euclidean Distance based Cost function	4.5019± 0.4455	1.8622± 0.4176	7.1065± 1.7463
KL Divergence Based Cost Function	4.4254± 0.5457	2.0707± 0.3122	7.7708± 1.8321

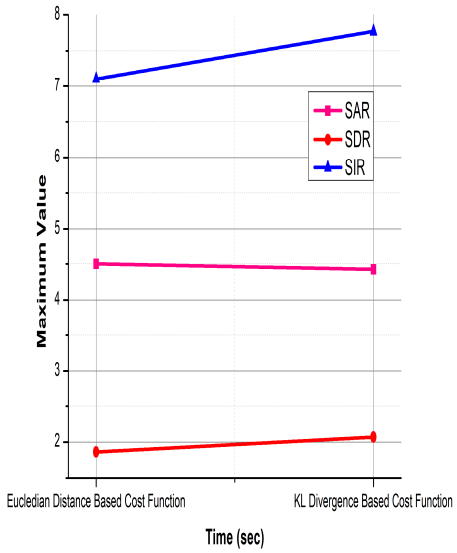

Fig. 10. Comparison of BSS performance using Basic NMF with KL Divergence and Euclidean distance-based Cost functions.

Table 2: Performance Evaluation of BSS using Different Types of NMF with Euclidean Distance based Cost function.

NMF Type	SAR	SDR	SIR
Basic NMF	4.5019 ±0.4455	1.8622±0.41 76	7.1065± 1.7463
Orthogonal NMF	4.2263± 0.3458	0.1786± 0.8005	3.4519± 0.8065
Sparse NMF	-3.9637± 0.6916	-7.9865± 1.4927	0.074± 2.2968

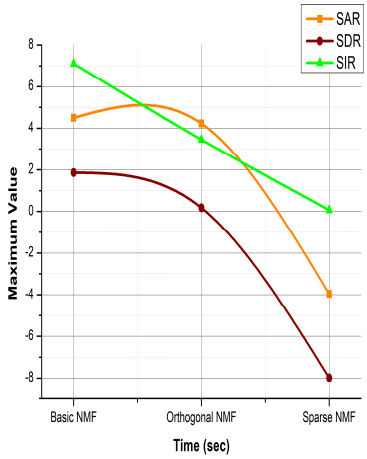

Fig. 11. Performance Evaluation of BSS using Different Types of NMF with Euclidean Distance based Cost function.

Table 3: Performance Evaluation of BSS using Different Types of NMF with KL Divergence based Cost function.

NMF Type	SAR	SDR	SIR
Basic NMF	4.4254±0.5457	2.0707±0.3122	7.7708±1.8321
Orthogonal NMF	4.0848±0.2194	1.3063±0.6130	6.3871±1.6988
Sparse NMF	4.0643±0.3388	1.6235±0.5330	7.2562±1.9359
Convolutive NMF	4.2530±0.0761	1.6578±0.2702	6.914±0.8287

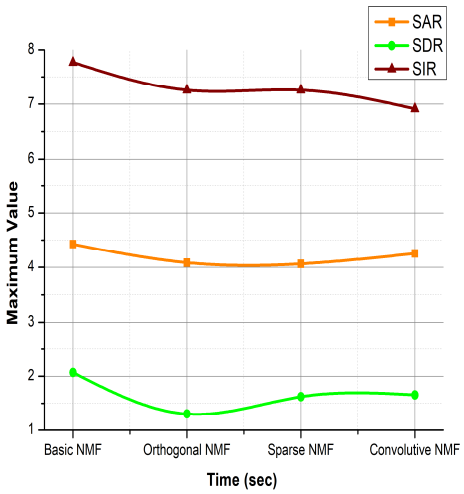

Fig. 12. Performance Evaluation of BSS using Different Types of NMF with KL Divergence based Cost function.

Table 4: Performance Evaluation after performing Speech enhancement techniques.

Speech enhancement method	NMF Algorithm	SAR	SDR	SIR
Adaptive noise cancellation	KL divergence	3.9047 ±0.554	2.0774 ±0.249	12.053 ±0.655
	Euclidean distance	3.9114 ±0.515	1.8843 ±0.362	8.2946 ±2.179

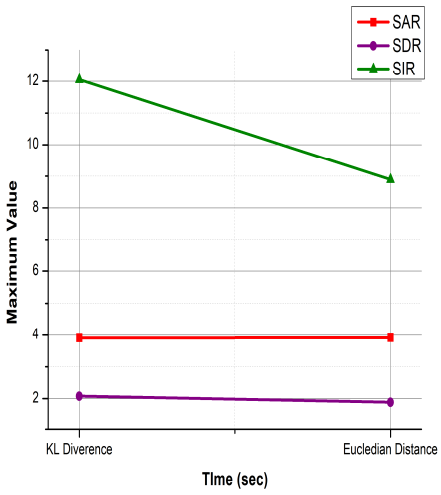

Fig. 13. Performance Evaluation after performing Speech enhancement techniques.

Table 5: BSS Performance Comparison with and without speech enhancement.

Method	SAR	SDR	SIR
GCC-NMF	4.4254 ±0.545	2.0707 ±0.312	7.7708 ±1.832
GCC-NMF with Spectral subtraction	3.3076 ±1.429	1.6387 ±0.583	9.8983 ±1.832
GCC-NMF with adaptive noise cancellation	3.9047 ±0.554	2.0774 ±0.249	12.053 ±0.655

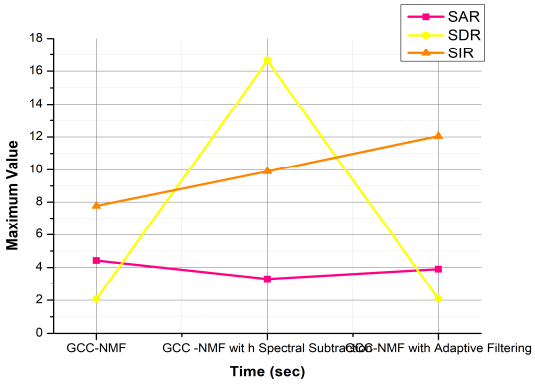

Fig. 14. BSS Performance Comparison with and without speech enhancement.

Summary. A new scheme is developed with the aid of the non-negative matrix factorization system for source separation in audio. Firstly, this method establishes the noise mixing paradigm and then incorporates the techniques of signal processing, where scattering is used for the study of characteristics. Wavelet filter banks have been built with the scattering method to use the optimum solution to obtain the filtered signal. Eventually, a channel isolation method is applied to reduce the resolution for each source to aid in audio demixing.

IV. CONCLUSION

This paper suggests a robust algorithm to differentiate the source signals from the weak combination. The method of isolation is implemented for using both Generalised Cross-Correlation and non-negative Matrix Factorisation—separation of the root. Speech amplification methods for adaptive filtering and spectral subtraction increase the separation efficiency of the source signals. Speech enhancement enhances the SIR efficiency of isolated outlets. The algorithm provides efficient results with some available approaches and it is proved its betterment with such methods.

V. FUTURE SCOPE

The future perspective is to improve the speech separation paradigms with spectral subtraction strategies. Secondly, to implement the simulation study in real-time event. By using the same methodology improve the pitch estimation. Reconstruction of speech signal using lip movement and facial expressions.

Conflict of Interest. There is no Conflict of Interest.

REFERENCES

[1]. Këpuska, V., & Bohouta, G. (2017). Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). *Int. J. Eng. Res. Appl.*, *7*(03), 20-24.

[2]. Kokpujie, K., Noma-Osaghae, E., John, S., & Jumbo, P. C. (2017, October). Automatic home appliance switching using speech recognition software and embedded system. In *2017 international conference on computing networking and informatics (ICCNI)* (pp. 1-4). IEEE.

[3]. Rebai, I., Ben Ayed, Y., Mahdi, W., & Lorré, J. P. (2017). Improving speech recognition using data augmentation and acoustic model fusion. *Procedia computer science*, *112*, 316-322.

[4]. Mustafa, M. K., Allen, T., & Appiah, K. (2019). A comparative review of dynamic neural networks and hidden Markov model methods for mobile on-device speech recognition. *Neural Computing and Applications*, *31*(2), 891-899.

[5]. Mukherjee, H., Obaidullah, S. M., Santosh, K. C., Phadikar, S., & Roy, K. (2018). Line spectral frequency-based features and extreme learning machine for voice activity detection from audio signal. *International Journal of Speech Technology*, *21*(4), 753-760.

[6]. Vincent, E., Watanabe, S., Nugraha, A. A., Barker, J., & Marxer, R. (2017). An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, *46*, 535-557.

[7]. Tabani, H., Arnau, J. M., Tubella, J., & Gonzalez, A. (2017). Performance analysis and optimization of automatic speech recognition. *IEEE Transactions on Multi-Scale Computing Systems*, *4*(4), 847-860.

[8]. Smit, P., Virpioja, S., & Kurimo, M. (2017, August). Improved Subword Modeling for WFST-Based Speech Recognition. In *INTERSPEECH* (pp. 2551-2555).

[9]. Errattahi, R., El Hannani, A., & Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, *128*, 32-37.

[10]. Hori, T., Cho, J., & Watanabe, S. (2018, December). End-to-end speech recognition with word-based RNN language models. In *2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 389-396). IEEE.

[11]. Haridas, A. V., Marimuthu, R., & Sivakumar, V. G. (2018). A critical review and analysis on techniques of speech recognition: The road ahead. *International Journal of Knowledge-Based and Intelligent Engineering Systems*, *22*(1), 39-57.

[12]. Czyzewski, A., Kostek, B., Bratoszewski, P., Kotus, J., & Szykulski, M. (2017). An audio-visual corpus for multimodal automatic speech recognition. *Journal of Intelligent Information Systems*, *49*(2), 167-192.

[13]. Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., & Mahjoub, M. A. (2017, May). A review on speech emotion recognition: Case of pedagogical interaction in classroom. In *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)* (pp. 1-7). IEEE.

[14]. Badshah, A. M., Ahmad, J., Rahim, N., & Baik, S. W. (2017, February). Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 international conference on platform technology and service (PlatCon)* (pp. 1-5). IEEE.

[15]. Zhang, Y., Chan, W., & Jaitly, N. (2017, March). Very deep convolutional networks for end-to-end speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4845-4849). IEEE.

[16]. Grozdić, Đ. T., Jovičić, S. T., & Subotić, M. (2017). Whispered speech recognition using deep denoising autoencoder. *Engineering Applications of Artificial Intelligence*, *59*, 15-22.

[17]. Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE Access*, *7*, 19143-19165.
[18]. Hori, T., Watanabe, S., & Hershey, J. R. (2017, December). Multi-level language modeling and decoding for open vocabulary end-to-end speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 287-293). IEEE.
[19]. Sarma, H., Saharia, N., & Sharma, U. (2017). Development and analysis of speech recognition systems for assamese language using HTK. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *17*(1), 1-14.
[20]. Huang, J., Sheffield, B., Lin, P., & Zeng, F. G. (2017). Electro-tactile stimulation enhances cochlear implant speech recognition in noise. *Scientific reports*, *7*(1), 1-5.
[21]. Bando, Y., Mimura, M., Itoyama, K., Yoshii, K., & Kawahara, T. (2018, April). Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 716-720). IEEE.
[22]. Tachioka, Y., Narita, T., Miura, I., Uramoto, T., Monta, N., Uenohara, S., & Le Roux, J. (2017, August). Coupled Initialization of Multi-Channel Non-Negative Matrix Factorization Based on Spatial and Spectral Information. In *INTERSPEECH* (pp. 2461-2465).
[23]. Shimada, K., Bando, Y., Mimura, M., Itoyama, K., Yoshii, K., & Kawahara, T. (2018, April). Unsupervised beamforming based on multichannel nonnegative matrix factorization for noisy speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5734-5738). IEEE.
[24]. Huang, S., Wang, H., Li, T., Li, T., & Xu, Z. (2018). Robust graph regularized nonnegative matrix factorization for clustering. *Data Mining and Knowledge Discovery*, *32*(2), 483-503.
[25]. Mohammed, S., & Tashev, I. (2017, March). A statistical approach to semi-supervised speech enhancement with low-order non-negative matrix factorization. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 546-550). IEEE.
[26]. Leglaive, S., Girin, L., & Horaud, R. (2019, May). Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 101-105). IEEE.
[27]. Leglaive, S., Badeau, R., & Richard, G. (2017, October). Separating time-frequency sources from time-domain convolutive mixtures using non-negative matrix factorization. In *2017 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)* (pp. 264-268). IEEE.
[28]. Erdogan, H., Hershey, J. R., Watanabe, S., & Le Roux, J. (2017). Deep recurrent networks for separation and recognition of single-channel speech in nonstationary background audio. In *New Era for Robust Speech Recognition* (pp. 165-186). Springer, Cham.
[29]. Chung, H., Plourde, E., & Champagne, B. (2017). Regularized non-negative matrix factorization with Gaussian mixtures and masking model for speech enhancement. *Speech Communication*, *87*, 18-30.
[30]. Lee, S., Han, D. K., & Ko, H. (2017). Single-channel speech enhancement method using reconstructive NMF with spectrotemporal speech presence probabilities. *Applied Acoustics*, *117*, 257-262.
[31]. J. Pooja, S. Rinesh., K. Logu (2020). Speaker Recognition System : *Test Engineering and Management*, 0193-4120 (82) .10420-10424.
[32]. Wood, S. U., Rouat, J., Dupont, S., & Pironkov, G. (2017). Blind speech separation and enhancement with GCC-NMF. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(4), 745-755.
[33]. Lee, R., Kang, M. S., Kim, B. H., Park, K. H., Lee, S. Q., & Park, H. M. (2020). Sound Source Localization Based on GCC-PHAT With Diffuseness Mask in Noisy and Reverberant Environments. *IEEE Access*, *8*, 7373-7382.
[34]. Sivasankaran, S., Vincent, E., & Fohr, D. (2020). SLOGD: Speaker Location Guided Deflation approach to speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6409-6413). IEEE.
[35]. Grondin, F., Tang, H., & Glass, J. (2020). Audio-Visual Calibration with Polynomial Regression for 2-D Projection Using SVD-PHAT. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4856-4860). IEEE.
[36]. Hu, J., Mo, Q., & Liu, Z. (2020). Multi-Source Classification: A DOA-Based Deep Learning Approach. In *2020 International Conference on Computer Engineering and Application (ICCEA)* (pp. 463-467). IEEE.
[37]. Murthy, B. N., Yegnanarayana, B., & Kadiri, S. R. (2020). Time delay estimation from mixed multispeaker speech signals using single frequency filtering. *Circuits, Systems, and Signal Processing*, *39*(4), 1988-2005.
[38]. Thakallapalli, S., Kadiri, S. R., & Gangashetty, S. V. (2020). Spectral Features derived from Single Frequency Filter for Multispeaker Localization. In *2020 National Conference on Communications (NCC)* (pp. 1-6). IEEE.
[39]. Zhang, T., Geng, Y., Sun, J., Jiao, C., & Ding, B. (2020). A Unified Speech Enhancement System Based on Neural Beamforming With Parabolic Reflector. *Applied Sciences*, *10*(7), 2218.
[40]. Hu, J., Mo, Q., & Liu, Z. (2020). Multi-Source Classification: A DOA-Based Deep Learning Approach. In *2020 International Conference on Computer Engineering and Application (ICCEA)* (pp. 463-467). IEEE.

How to cite this article: Pradeep, R., Kanimozhi, R., Prajitha, C. and Rinesh, S. (2020). Non- Negative Matrix Factorization based Blind Source Separation and Source Enhancement using Generalised Cross Correlation. *International Journal on Emerging Technologies*, *11*(3): 927–935.