Blind Source Separation based on Semblance Beamforming

Alexandre Miccheleti Lucena, Kenji Nose-Filho, Ricardo Suyama

***Abstract*— The source separation task has been tackled from different approaches, including beamforming. The idea of exploiting geometric information of the sensors may contribute in convolutive mixture separation context, usually assumed in real scenario. This work proposes a beamformer algorithm for source separation based on the semblance coherence function. The performance of the algorithm for artificially mixed signals is evaluated using a objective intelligibility metric throughout Monte Carlo simulations in two scenarios and different SNR levels. Results are compared with classic techniques: GSS and Delay and Sum, where the proposed algorithm achieves the best performance under no influence of additive noise.**

***Keywords*— source separation, beamforming, semblance, source cancellation, time difference of arrival (TDOA), convolutive mixtures.**

I. Introduction

The signal separation or source separation task can be a common necessity in different practical applications. There is an increasing interest of achieving this goal in the speech recognition context (e.g. hands-free communication, rescue scenarios, etc.) as the acquired signals are often corrupted by noise from the environment or other unwanted sources.

We can cite two main categories of signal processing techniques that focus on solving the source separation problem: blind source separation (BSS) and beamforming. As discussed in [1], both techniques may sometimes have similar objectives but different approaches. While BSS usually takes advantage of a prior knowledge of sources characteristics (e.g. second-order statistics), beamforming techniques exploit sensors geometry information, but both approaches proved their capability in achieving satisfactory separation, or even shown to be equivalent in some convolutive mixture separation context [2]. As a way of improving the BSS performance, there are initiatives as the *Geometric Source Separation* (GSS) in [3] that combines both methods in order to achieve separation.

For speech and audio signal separation, the acoustic modelling involved on the acquisition process is a important practical consideration. Many beamforming techniques [4]-[5] rely on time difference of arrival (TDOA) as a way to estimate the direction of arrival (DOA) of a sound source. Recent works have presented a new semblance based TDOA algorithm that showed its potential in source localization and speech enhancement contexts [6]-[7].

The objective of this paper is the development of a beamformer for source separation based on the semblance TDOA algorithm. Our approach consists in finding the direction of arrival of a given source and then subtracting it from the mixture by using a least squares filter or a least absolute deviation filter. The algorithm is tested in a simulation environment, with artificially convolved mixed sound sources, and has its performance evaluated in comparison to classic approaches, the Delay and Sum beamformer and the GSS algorithm.

II. Problem Definition and Methods

In a real scenario, the observed signal $x_i(n)$, for N sources and M microphone recordings, can be written in terms of a convolutive mixture discrete model as

$$x_i\left(n\right) = \sum_{j=1}^{N}\sum_{p=1}^{P} a_{ij}(p) s_j\left(n - \tau_{ipj}\right), \left(i = 1, \cdots, M\right), \quad (1)$$

where s_j is the source signal from source j, a_{ij} are the filter coefficients, and P the number of paths from sources to microphone, causing its respective τ_{ipj} delay. For an anechoic environment, i.e., without multipath propagation, (1) becomes

$$x_i\left(n\right) = \sum_{j=1}^{N} a_{ij} s_j\left(n - \tau_{ij}\right), \left(i = 1, \cdots, M\right), \quad (2)$$

being τ_{ij} the delay caused by line of sight propagation distance.

One way to accomplish separation in such scenario (anechoic mixtures) is to obtain the unmixing filter weights, which are applied to each microphone signal. The resulting filtered signals are the estimated sources. The described process can be modeled as

$$y_k\left(n\right) = \sum_{i=1}^{M} w_{ki} x_i\left(n - \hat{\tau}_{ki}\right), \left(k = 1, \cdots, N\right), \quad (3)$$

where w_{ki} are the unmixing filter weights, $\hat{\tau}_{ki}$ the unmixing delays, and y_k the estimated source.

A. Delay and Sum beamformer

One of the simplest ways to perform source separation is given by the Delay and Sum beamformer. From the source direction, it is possible to estimate the delays $\hat{\tau}_{ki}$ such that $\tau_{ij} + \hat{\tau}_{ki} = \tau_k$ are the same for $j = k$, so that equation (3), for unitary weigths $w_{ki} = 1$, becomes

$$y_k\left(n\right) = \sum_{i=1}^{M} a_{ik} s_k\left(n - \tau_k\right) + \sum_{i=1}^{M}\sum_{j\neq k} a_{ij} s_j\left(n - \left(\tau_{ij} + \hat{\tau}_{ki}\right)\right). \quad (4)$$

Alexandre Miccheleti Lucena, Kenji Nose-Filho and Ricardo Suyama are with the Engineering, Modeling and Applied Social Sciences Center, Federal University of ABC, Santo André, SP, Brazil, E-mails: alexandre.lucena@ufabc.edu.br, kenji.nose@ufabc.edu.br, ricardo.suyama@ufabc.edu.br. This study was financed in part by the Coordenação de Aperfeiçoeamento de Nível Superior - Brasil (CAPES) - Finance Code 001 - and the National Council for Scientific and Technological Development (CNPq)

Intuitively, the Delay and Sum beamformer tries to promote a constructive interference of the aligned wavefronts (first term on the right-hand side of equation (4)) and a destructive interference from the non aligned wavefronts (second term on the right-hand side of equation (4)).

B. Geometric Source Separation

The Geometric Source Separation method proposed in [3] is based on the context of blind source separation of convolutive mixtures with geometric constraints. Considering the Short Time Fourier Transform (STFT), the separated signals for discrete frequency ω at the time frame l are given by

$$\mathbf{Y}(\omega,l) = \mathbf{W}(\omega,l)\mathbf{X}(\omega,l), \tag{5}$$

where $\mathbf{X}(\omega,l)$ is the STFT of the mixtures and $\mathbf{W}(\omega,l)$ is the unmixing matrix for each frequency bin ω and time frame l. For simplicity, ω and l will be supressed from notation. The GSS algorithm aims to minimize the following cost functions:

$$J_1(\mathbf{W}) = \|\mathbf{R_{YY}} - \text{diag}[\mathbf{R_{YY}}]\|^2, \tag{6}$$

$$J_2(\mathbf{W}) = \|\mathbf{WA} - \mathbf{I}\|^2, \tag{7}$$

where $\|\cdot\|$ stands for the matrix norm, given by $\|\mathbf{X}\| = \sqrt{\text{trace}[\mathbf{XX}^H]}$, and $\mathbf{R_{YY}}$ is the correlation matrix of $\mathbf{Y}(\omega,l)$. Each element of matrix $\mathbf{A}$ is given by $A_{ij}(\omega) = e^{-j\omega\tau_{ij}}$. (6) promotes the decorrelation of the estimated signals and (7) is a geometric constraint imposed by the direction of arrival of each source. It is interesting to notice that, for $\mathbf{W} = \mathbf{A}^H$, the GSS is equivalent to the Delay and Sum beamformer [3].

III. Semblance based Beamformer (SBB)

In this paper, we propose a new way to perform source separation following the same idea explored by the DS beamformer. However, instead of promoting a constructive interference of the aligned wavefronts we are going in the opposite direction, trying to promote a destructive interference of the aligned wavefronts, being able to remove or attenuate the influence of the source of a given direction. This is a similar technique applied in seismic reflection for the attenuation of the surface-related multiple reflections [8], [9].

First, let us consider a linear array with two microphones ($M = 2$), spaced with a distance equal to d. Then, if we consider a plane wavefront, with an angle of incidence equal to θ_k, the time difference between the signals received by the microphones is equal to $\hat{\tau}_k = \frac{d\sin(\theta_k)}{c}$, where c is the speed of sound in air. By applying this delay to the signals received by the microphones, we have

$$\begin{aligned} \hat{x}_{k1}(n) &= x_1(n), \\ \hat{x}_{k2}(n) &= x_2(n - \hat{\tau}_k), \end{aligned} \tag{8}$$

as being the signals with the aligned wavefronts from the k-th source, considering $x_1(n)$ the recording of the microphone to the left and $x_2(n)$ the recording of the microphone to the right.

To remove or attenuate the influence of the k-th source, we propose a simple filtering scheme

$$y_k(n) = \hat{x}_{k1}(n) - w_k\hat{x}_{k2}(n). \tag{9}$$

Substituting equation (2) into (9) yields (10) (located at the top of the next page). By finding $w_k = \frac{a_{1k}}{a_{2k}}$, equation (10) becomes

$$y_k(n) = \sum_{j\neq k}^{N} a_{1j}s_j(n-\tau_{1j}) - w_k a_{2j} s_j(n - \tau_{2j} - \hat{\tau}_k). \tag{11}$$

By assuming that there is no destructive interference from the other wavefronts, the signal $y_k(n)$ can be seen as the sum of filtered versions of all the other sources, except for $s_k(n)$.

For the simple case with only two active sources, i.e., $N = 2$, with different angles of incidence, we have, for $w_1 = \frac{a_{11}}{a_{21}}$ and $w_2 = \frac{a_{12}}{a_{22}}$

$$\begin{aligned} y_1(n) &= a_{12}s_2(n-\tau_{12}) - \tfrac{a_{11}}{a_{21}}a_{22}s_2(n-\tau_{22}-\hat{\tau}_1), \\ y_2(n) &= a_{11}s_1(n-\tau_{11}) - \tfrac{a_{12}}{a_{22}}a_{21}s_1(n-\tau_{21}-\hat{\tau}_2). \end{aligned} \tag{12}$$

So $y_1(n)$ can be seen as a filtered version of $s_2(n)$ and $y_2(n)$ can be seen as a filtered version of $s_1(n)$. To avoid confusion, in the Section Results we will call $y_1(n)$ as being the estimate for $s_1(n)$ and $y_2(n)$ as being the estimate for $s_2(n)$.

To estimate the direction of arrival and, consequently, the value of $\hat{\tau}_k$ we use the same algorithm as in [6] but in the frequency domain, which enables us to work at low sampling rates. In [6] the authors propose a semblance coherence function based TDOA algorithm, that is a widely used measurement in multichannel data in seismic signal processing, to measure the level of similarity of signals [10]. In the frequency domain, due to the Parseval's Theorem, the semblance cost function becomes

$$Z_d = \frac{\sum_\omega |\sum_i \hat{X}_i(\omega)|^2}{M\sum_\omega \sum_i |\hat{X}_i(\omega)|^2}. \tag{13}$$

For $N > 2$, by direct applying the proposed method we can remove one of the sources from the recordings, by knowing its direction.

In the following, to estimate w_k, we propose two methods: A filter based on the least squares, and another one on the least absolute deviations.

A. Least Squares Filter

The first proposed method to estimate w_k, based on a minimum energy assumption, is to minimize the least squares error criterion (or L_2 norm) as in

$$w_k = \underset{w_k}{\text{argmin}} \quad \|\hat{x}_{k1}(n) - w_k\hat{x}_{k2}(n)\|_2^2, \tag{14}$$

$$w_k = \frac{\sum_n \hat{x}_{k1}(n)\,\hat{x}_{k2}(n)}{\sum_n \hat{x}_{k2}^2(n)}. \tag{15}$$

B. Least Absolute Deviation Filter

The second method to estimate w_k, based on a sparsity assumption, is to minimize the least absolute deviation (or L_1 norm) as in

$$w_k = \underset{w_k}{\text{argmin}} \quad \|\hat{x}_{k1}(n) - w_k\hat{x}_{k2}(n)\|_1, \tag{16}$$

$$\begin{aligned} &y_k\left(n\right) = x_1\left(n\right) - w_k x_2\left(n - \hat{\tau}_k\right), \\ &y_k\left(n\right) = \textstyle\sum_{j=1}^{N} a_{1j} s_j\left(n - \tau_{1j}\right) - w_k \sum_{j=1}^{N} a_{2j} s_j\left(n - \tau_{2j} - \hat{\tau}_k\right), \\ &y_k\left(n\right) = \left(a_{1k} - w_k a_{2k}\right) s_k(n - \tau_{1k}) + \textstyle\sum_{j\neq k}^{N} a_{1j} s_j\left(n - \tau_{1j}\right) - w_k a_{2j} s_j\left(n - \tau_{2j} - \hat{\tau}_k\right). \end{aligned} \tag{10}$$

where w_k can be searched by an iterative gradient descent procedure. With a random initialization of w_k^0, its update for each iteration m is given by

$$w_k^{m+1} = w_k^m + \mu \sum_n \operatorname{sgn}\left(\hat{x}_{k1}(n) - w_k^m \hat{x}_{k2}\left(n\right)\right) \hat{x}_{k2}(n). \tag{17}$$

IV. Simulation Setup

The audio files of two different speakers with 5 seconds duration used in simulations as sources, were extracted from the [BRSpeechCorpus][1] dataset. In order to compare the performance of the proposed methods, the GSS and Delay and Sum algorithms were also implemented in the same conditions for posterior evaluation. The source signals were mixed artificially assuming an anechoic environment and far-field positioning. Before the mixing process both sources were normalized by its standard deviation. The sources amplitudes were multiplied by a value obtained through normalization of the radius (distance between source and sensor) by the larger radius. This value was multiplied by a random variable with distribution $\mathcal{N}(1, 0.25)$, to attribute different random amplitude to the sources. The proposed method was tested in a simulation environment in two different scenarios, both simulating two sources (N=2) and two microphones (M=2).

1) Distant Sources: The position x and y of source s_1 was generated by sampling two independent random variables with gaussian distribution of mean 1500cm with a standard deviation of 150cm. Source s_2 position coordinates x and y were generated by two independent random variables with gaussian distribution of mean -1200cm and 1800cm respectively, with standard deviation of 150cm.

2) Near Sources: The position x and y of source s_1 was generated by sampling two independent random variables with gaussian distribution of mean 0cm and 1500cm respectively, with a standard deviation of 150cm. Source s_2 position coordinates x and y were generated by two independent random variables with gaussian distribution of mean -900cm and 1800cm respectively, with standard deviation of 150cm.

In each scenario, the algorithm was also tested for different levels of interfering noise on the sensors. For this purpose, artificially generated white Gaussian noise was added to the mixed signals to match the desired level of signal-to-noise ratio (SNR).

V. Results and Discussion

As described in simulation setup, the mixed signals were obtained through an artificial mixing process and different SNR levels. The algorithms for the proposed methods, Delay and Sum, and GSS were implemented for scenarios 1 and 2 under the same conditions.

The performance evaluation was made by calculating the short-time objective intelligibility measure (STOI) [11], of the estimated signals for each algorithm. The STOI is an objective measure specifically developed for speech signal evaluation, and it is a function of the clean and degraded speech, measuring an intelligibility score between 0 and 1. It can be interpreted as the percentage of the original signal that can be comprehended in the estimated signal. For the STOI calculation step, all the estimated signals had their delay corrected, correlating it to its respective original source.

The STOI was calculated for every estimated signal $y_k(n)$ with respect to all source signals $s_k(n)$. When the index k in $y_k(n)$ and $s_k(n)$ is equal, ideally it is wanted a STOI score closer to 1. However, when the index k is different, a score closer to 0 is desired, meaning that the original signal is not present in the estimated signal. This is an important metric not only to estimate the quality of the estimated source signal, but to determine how well it was subtracted from other estimates.

The algorithms were tested in a noiseless scenario and 4 different SNR conditions: 30dB, 25dB, 20dB, 15dB. In order to provide more consistent results, each experiment was repeated 100 times (Monte Carlo method), for each one of the SNR conditions the mean STOI was calculated. Also, the STOI values for each SNR condition provide a way to understand how well the algorithms perform over noisy sensors.

A. *Scenario 1: Distant Sources*

Figure 1 presents the simulation results obtained for scenario 1 (*Distant Sources*). Figures from 1.(a) to 1.(d) show the mean STOI values for each algorithm and SNR levels. The STOI was calculated over the delay corrected mixed signals $x_k(n)$ with no further treatment, as a way to compare each algorithm improvement over it.

In Fig. 1.(a) it can be seen that the proposed method outperforms the GSS and Delay and Sum algorithms when noise is absent, with a slightly better result for the L_2 norm based method, with a STOI of 0.92. For noisy conditions, the proposed method outperforms the Delay and Sum algorithm, almost reaching GSS, presenting lower STOI values but similar performance. Figures 1.(b) and 1.(c) show that the proposed method significantly reduced the influence of the undesired sources. Figures 1.(b) indicates a similar performance to the GSS, with better results for a SNR of 20dB and 15dB. Fig. 1.(d) shows that the proposed method has the best performance without the addition of noise, with a STOI of 0.96 for the L_1 norm based method. For noisy conditions, the proposed method outperforms the Delay and Sum algorithm for SNR of 20dB, 25dB and 30dB, and improves the STOI considerably.

Figures 1.(e) and 1.(f), show the average error associated with the angle estimation in scenario 1 of the semblance based TDOA algorithm for the angle of incidence of source 1 and source 2, respectively. For both sources, it can be seen that the

[1] Dataset available at: http://lasp.ime.eb.br/index.php?vPage=downloads

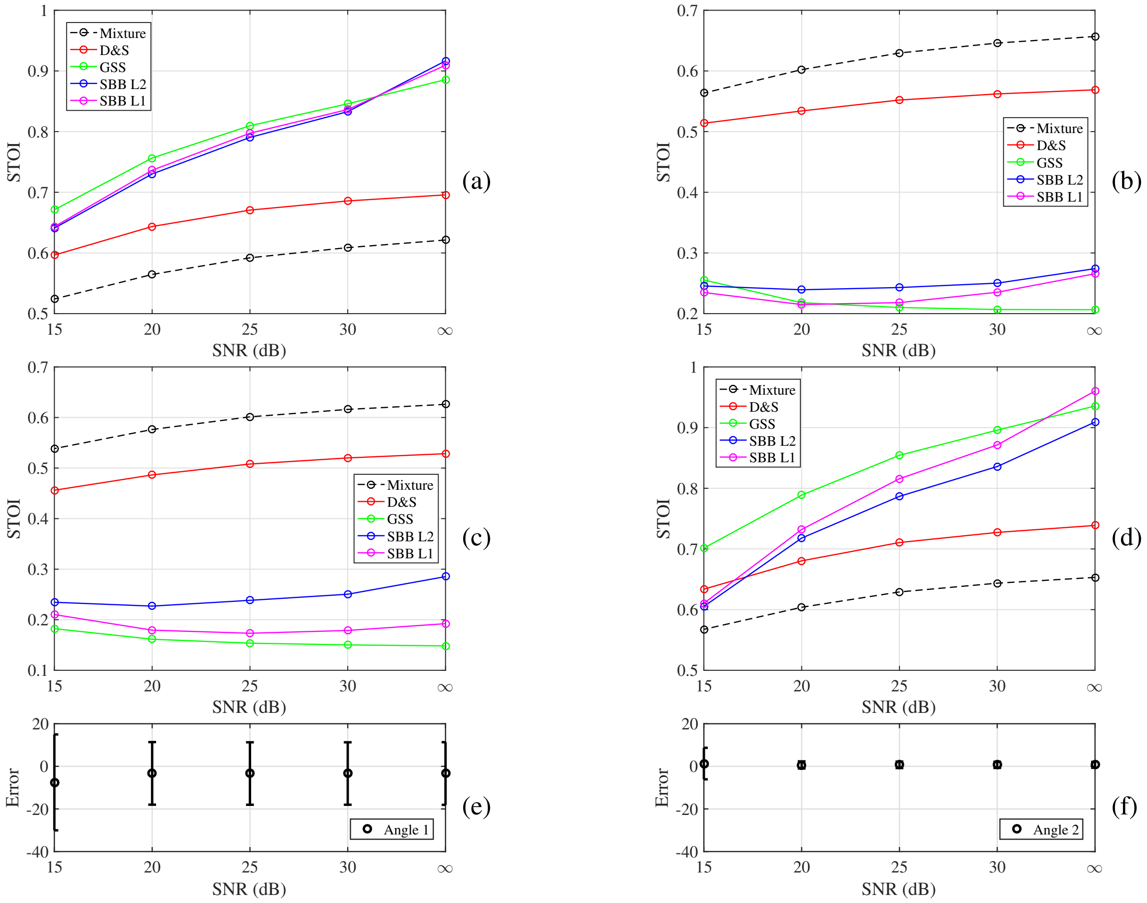

Fig. 1. Comparison of STOI values of (a) $y_1(n)$ with respect to $s_1(n)$, (b) $y_1(n)$ with respect to $s_2(n)$, (c) $y_2(n)$ with respect to $s_1(n)$ and (d) $y_2(n)$ with respect to $s_2(n)$ for different SNR for Scenario 1. (e) and (f) illustrates the mean and standard deviation of the errors (in degrees) of the semblance based TDOA algorithm, for source 1 and 2, respectively. It is important to note that the Delay and Sum and the GSS algorithms for the separation of source 1 depend on the estimation of the DOA of source 1, while the proposed algorithm depends on the estimation of the DOA of source 2 to enhance the Source 1 Signal.

angle estimation error increases as the SNR level decreases, especially for angle 1. Since all algorithms depends on the angle estimation, their overall performance reduction can be associated with the increase in the angle estimation error. However, for the SBB algorithm, the enhancement of the signal source 1 depends on the estimation of the DOA of source 2, and vice versa. For this reason, the increase in the estimation error of angle 1 in Fig. 1.(e) impacts the proposed algorithm performance observed in Fig. 1.(d).

B. Scenario 2: Near Sources

Fig. 2 presents the simulation results obtained for scenario 2 (*Near Sources*). Figures from 2.(a) to 2.(d) shows the mean STOI values for each algorithm and SNR levels. As for scenario 1, the STOI was calculated over the delay corrected mixed signals $x_k(n)$, for comparison purposes.

Without noise addition, the proposed method surpass the performance of the other algorithms as seen in Fig. 2(a), with a score of 0.94 for the L_2 norm based method. In the presence of noise, the proposed method has an equivalent performance to the GSS algorithm for SNR levels of 30dB and 25dB. In Figures 2(b) and 2(c) the best subtraction results are from the L_2 norm based SBB for both sources, while L_1 norm based method equate it's performance as the noise level increases. In Fig. 2(d), with a STOI of 0.93 the L_2 norm based SBB presents the best result in the absence of noise, and GSS has the best performance in noisy conditions.

Figures 2.(e) and 2.(f), show the average error associated with the angle estimation in scenario 2 of the semblance based TDOA algorithm for the angle of incidence of source 1 and source 2, respectively. For near sources (scenario 2), the angles of incidence of the sources over the sensors have closer values, and the overall angle estimation error increases in relation to scenario 1. As seen in scenario 1, once again in scenario 2, there is an impact on overall performance of the algorithms with the increase of the noise, and consequently there is an increase in the angle estimation error. Since SBB algorithm estimation of a source depends on the DOA of the opposite source, the decrease of performance of proposed algorithm in Fig. 2.(d) can be justified by the increase of the angle estimation error in 2.(e).

VI. Conclusion

In this work, we proposed a new beamformer algorithm based on the semblance coherence function for source separation, by exploring its capability of estimating the DOA of the sources followed by a filtering step. In the case of absence of additive noise on the mixed signals, the proposed method based on

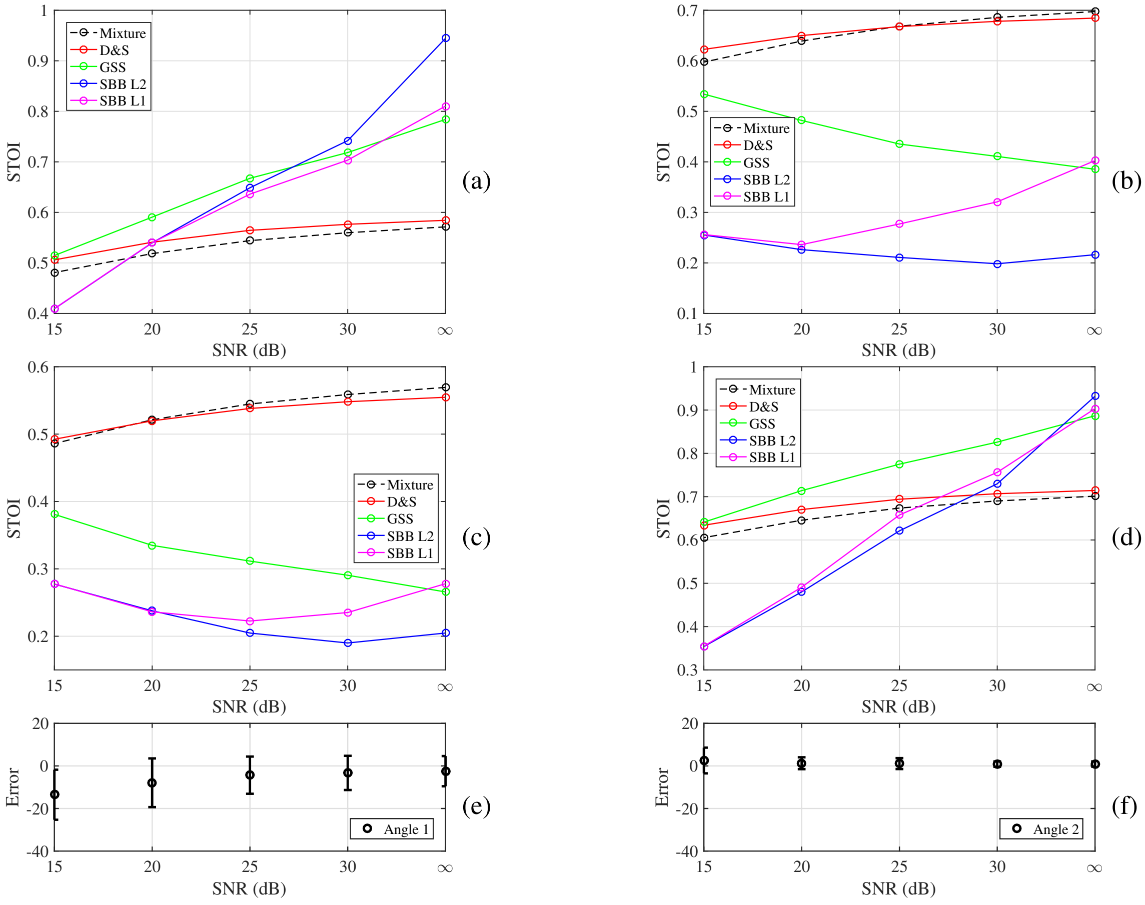

Fig. 2. Comparison of STOI values of (a) $y_1(n)$ with respect to $s_1(n)$, (b) $y_1(n)$ with respect to $s_2(n)$, (c) $y_2(n)$ with respect to $s_1(n)$ and (d) $y_2(n)$ with respect to $s_2(n)$ for different SNR for Scenario 2. (e) and (f) illustrates the mean and standard deviation of the errors (in degrees) of the semblance based TDOA algorithm, for source 1 and 2, respectively. It is important to note that the Delay and Sum and the GSS algorithms for the separation of source 1 depend on the estimation of the DOA of source 1, while the proposed algorithm depends on the estimation of the DOA of source 2 to enhance the Source 1 Signal.

L_2 norm had better overall average performance in recovering sources than other algorithms in both scenarios. In Scenario 1, the L_1 norm based method presented a superior capability in cancelling sources than the L_2 based method, whereas for Scenario 2 the L_2 based filtering had the lowest average STOI for all SNR levels in the source cancelling task. The advantage of the proposed method showed to be its lower complexity and fast execution time. It was capable of achieving better results than other methods for mixtures without noise, or even similar results and fastest execution time than GSS for different SNR levels. As future work, the perspective is to test the algorithm for real mixture recordings, and also evaluate its performance in estimating $N > 2$ sources.

References

[1] H. Adel, M. Souad, A. Alaqeeli, and A. Hamid, “Beamforming techniques for multichannel audio signal separation,” *arXiv preprint arXiv:1212.6080*, 2012.

[2] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, “Equivalence between frequency-domain blind source separation and frequency-domain adaptive beamforming for convolutive mixtures,” *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 11, p. 198923, 2003.

[3] L. C. Parra and C. V. Alvino, “Geometric source separation: Merging convolutive source separation with geometric beamforming,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[4] X. Anguera, C. Wooters, and J. Hernando, “Acoustic beamforming for speaker diarization of meetings,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[5] J. Bitzer and K. U. Simmer, “Superdirective microphone arrays,” in *Microphone arrays*. Springer, 2001, pp. 19–38.

[6] G. Aldeia, A. Crispim, G. Barreto, K. Alves, H. Ferreira, and K. Nose-Filho, “A semblance based tdoa algorithm for sound source localization.” XXXVII Simpósio Brasileiro de Telecomunicações e Processamento de Sinais, 10 2019.

[7] T. Spadini, G. S. I. Aldeia, G. Barreto, K. Alves, H. Ferreira, R. Suyama, and K. Nose-Filho, “On the application of segan for the attenuation of the ego-noise in the speech sound source localization problem,” in *2019 Workshop on Communication Networks and Power Systems (WCNPS)*. IEEE, 2019, pp. 1–4.

[8] D. J. Verschuur, *Seismic multiple removal techniques: past, present and future*. EAGE, 2006.

[9] D. Donno, “Improving multiple removal using least-squares dip filters and independent component analysis,” *Geophysics*, vol. 76, no. 5, pp. V91–V104, 2011. [Online]. Available: http://library.seg.org/doi/abs/10.1190/geo2010-0332.1

[10] N. S. Neidell and M. T. Taner, “Semblance and other coherency measures for multichannel data,” *GEOPHYSICS*, vol. 36, no. 3, pp. 482–497, 1971. [Online]. Available: https://doi.org/10.1190/1.1440186

[11] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.