Audiovisual Speech Separation based on Independent Vector Analysis using a Visual Voice Activity Detector

Pierre Narvor, Bertrand Rivet, and Christian Jutten

Univ. Grenoble Alpes, Gipsa-Lab, F-38000 Grenoble, France
CNRS, GIPSA-Lab, F-38000 Grenoble, France
{pierre.narvor,bertrand.rivet,christian.jutten}@gipsa-lab.fr

Abstract. In this paper, we present a way of improving the Independent Vector Analysis in the context of blind separation of convolutive mixtures of speech signals. The periods of activity and inactivity of one or more speech signals are first detected using a binary visual voice activity detector based on lip movements and then fed into a modified Independent Vector Analysis algorithm to achieve the separation. Presented results show that this approach improves separation and identification of sources in a determined case with a higher convergence rate, and is also able to enhance a specific source in an underdetermined mixture.

Keywords: Audiovisual Speech Separation, Convolutive Mixture, Blind Source Separation, Visual Voice Activity Detector, Independent Vector Analysis, Multimodality.

1 Introduction

The problem of extracting a speech signal of interest from a mixture of sounds in a natural reverberant environment is still a difficult task. This problem, well known as the cocktail-party problem [3], has been heavily investigated within the field of Convolutive Blind Source Separation (CBSS) in the past decades [7]. The Independent Vector Analysis (IVA) framework introduced in [5] and similarly in [4] has been proposed as a possible way of achieving such a separation. Indeed, the CBSS can be performed in the frequency space. Each frequency bin of the Discrete Short-Term Fourier Transform (D-STFT) of the observed signal is a linear instantaneous mixture of the D-STFT of the source signals. Therefore, the separation can be carried out at each frequency bin using Independent Component Analysis (ICA). However, because of the permutation ambiguity inherent to blind separation of signals, a random permutation between frequency

This work have been partly supported by the ERC project CHESS: 2012-ERC-AdG-320684

bins occurs during the separation process. A post-processing step is thus needed to reassociate the frequency bins to the proper sources, as in [9]. On the contrary, the IVA is able to perform a joint ICA of frequency bins, allowing to keep a coherence between those. Unfortunately, even if each estimated frequency bin is associated to the right estimated source, a global permutation indeterminacy still remains between the sources. This could be a problem in a case where there are less sources of interest than the total number of sources. Identifying the right ones might be a real challenge without further information.

In the context of separation or extraction of speech signals, these further information can be given by a video of the speaker's face. Indeed, the information carried by a video of the speaker's face is strongly related to the speech signal itself [11], but usually independent from the remaining sounds of the scene. For a recent overview of the field of audiovisual speech source separation, see [8]. In this paper, we propose to use a Visual Voice Activity Detector (V-VAD) to get the periods when the speech signal is actually active. Then, this binary information is included into an IVA algorithm to perform the separation. Presented results show that the estimated sources are associated to the right estimated activity after separation. The extraction is also faster and the quality is higher than when the estimated activity is not used. Moreover, we show that this method can also be used to enhance a specific source of interest in an underdetermined mixture. The method, designated as AV-IVA in the following, is compared to a reference IVA algorithm in which no other information than the audio is used. The method is also compared to an IVA where the actual information of speech activity is given by an oracle (O-IVA). The O-IVA gives us the highest performance bound that can be expected from this method.

Mathematical notations and mixing and separation models are defined in Section 2. The IVA algorithm is shortly described in Section 3, before a detailed presentation of our contribution. Experiments descriptions, numerical results and a discussion can be found in Section 4.

2 Mathematical Preliminaries

2.1 Notations

For now, only the determined case is considered. The number of sources N is the same than the number of microphones. Since the separation is processed in the frequency domain, the audio signals are represented by their D-STFT. The D-STFTs are processed over T frames of size *2(K-1)* time samples. The signals are real in the time domain, so the first K points of the Discrete Fourier Transform (DFT) are sufficient to represent a frame in the frequency domain. Finally, a complete set of audio data is represented by a 3D array of $N \times T \times K$ complex numbers. The arrays associated to the sources, the observations, and the estimated sources are denoted $\boldsymbol{S} \in \mathbb{C}^{N\times T\times K}$, $\boldsymbol{X} \in \mathbb{C}^{N\times T\times K}$ and $\boldsymbol{Y} \in \mathbb{C}^{N\times T\times K}$, respectively. In the latter, we will work on subsets of these arrays. An element of one array is denoted x_{ntk}. Vectors taken along the first, second and third dimensions are denoted $\boldsymbol{x}_{:tk}$, $\boldsymbol{x}_{n:k}$ and $\boldsymbol{x}_{nt:}$, respectively. Unless specified otherwise, these vectors are considered as column vectors. The matrices $\boldsymbol{X}_{n::} \in \mathbb{C}^{T\times K}$ and $\boldsymbol{X}_{::k} \in \mathbb{C}^{N\times T}$ are slices of a 3D array respectively orthogonal to the first and third dimensions. $\boldsymbol{X}_{n::}$ is the D-STFT of the n-th observation and $\boldsymbol{X}_{::k}$ is the k-th set of frequency bins where each line is the k-th bin of the D-STFT of an observation. The determinant of a matrix is denoted det, the Hermitian transposition operator is denoted $'$ and the identity matrix is denoted $\boldsymbol{I}$. To select the k-th element in a vector, we use a dot product with the k-th column vector of the canonical basis, denoted $\boldsymbol{e}_k$. The probability density function (pdf) of a uni/multivariate random variable is denoted $P_{\boldsymbol{y}_n}(\boldsymbol{y}_{nt:};\boldsymbol{\theta})$, with $\boldsymbol{\theta}$ a set of parameters defining the pdf.

2.2 Mixing and Separation Models

The audio mixing model is the classical mixing model of CBSS expressed in the frequency domain [7]. Each microphone picks a sum of several source contributions with an eventual additive noise. The relationship between the i-th source and its contribution to the j-th microphone is modeled as a convolution. This convolution becomes a product in the frequency domain if the analysis window is long compared to the impulse response of the filter. Under these assumptions the mixing model can be written as follows:

$$\boldsymbol{X}_{::k} = \boldsymbol{A}_{::k}\boldsymbol{S}_{::k} \qquad \forall k = 1\dots K. \tag{1}$$

$\boldsymbol{A} \in \mathbb{C}^{N\times N\times K}$ is a 3D array in which the DFTs of the mixing filters are stored along the third dimension. These impulse responses are *2(K-1)* points long but their DFTs are represented by K points since those are real. The separation process which estimates the sources is represented by $\boldsymbol{W} \in \mathbb{C}^{N\times N\times K}$ and is defined in a similar fashion to (1): $\boldsymbol{Y}_{::k} = \boldsymbol{W}_{::k}\boldsymbol{X}_{::k}, \forall k = 1\dots K$.

3 Method

Our contribution is based on the IVA algorithm described in [2]. In this section, we first recall this method. For the sake of simplicity we shall refer to it in the following as the reference IVA algorithm. The proposed improvement is presented in the subsection 3.3.

3.1 Cost function and learning algorithm

This algorithm is based on a Maximum Likelihood Estimation. The time samples are considered independent and identically distributed, and the sources are considered mutually independent. The cost function to minimize is then:

$$\boldsymbol{C}_{IVA} = -\frac{1}{T}\log P_{\boldsymbol{X}}(\boldsymbol{X}) = -\frac{1}{T}\sum_{n=1}^{N}\sum_{t=1}^{T}\log P_{\boldsymbol{y}_n}(\boldsymbol{y}_{nt:};\boldsymbol{\theta}) - \sum_{k=1}^{K}\log|\det \boldsymbol{W}_{::k}|. \tag{2}$$

The optimization is carried out with a natural gradient learning rule [1]. The derivation of the cost function is straightforward and starts with the computation of the regular gradient:

$$\frac{\partial \boldsymbol{C}_{IVA}}{\partial w_{n_1 n_2 k}} = -\frac{1}{T}\sum_{t=1}^{T} \frac{\partial \log P_{\boldsymbol{y}_{n_1}}(\boldsymbol{y}_{n_1 t:};\boldsymbol{\theta})'}{\partial \boldsymbol{y}_{n_1 t:}} \frac{\partial \boldsymbol{x}_{n_2 t:}}{\partial w_{n_1 n_2 k}} - \boldsymbol{e}'_{n_2}(\boldsymbol{W}^{-1}_{::k})\boldsymbol{e}_{n_1}. \tag{3}$$

Denoting $\boldsymbol{\Phi} \in \mathbb{C}^{N\times T\times K}$ by $\phi_{ntk} = (\partial \log P_{\boldsymbol{y}_n}(\boldsymbol{y}_{nt:};\boldsymbol{\theta})/\partial \boldsymbol{y}_{nt:})\boldsymbol{e}_k$, the gradient can be expressed in a more compact way: $\nabla_{\boldsymbol{W}_{::k}} \boldsymbol{C}_{IVA} = -\frac{1}{T}\boldsymbol{\Phi}_{::k}\boldsymbol{X}'_{::k} - (\boldsymbol{W}^{-1}_{::k})'$. The natural gradient, defined as $\widetilde{\nabla}_{\boldsymbol{W}_{::k}} \boldsymbol{C}_{IVA} = \nabla_{\boldsymbol{W}_{::k}} \boldsymbol{C}_{IVA}.\boldsymbol{W}'_{::k}$, can then be computed as follows:

$$\widetilde{\nabla}_{\boldsymbol{W}_{::k}} \boldsymbol{C}_{IVA} = -\frac{1}{T}\boldsymbol{\Phi}_{::k}\boldsymbol{Y}'_{::k} - \boldsymbol{I}. \tag{4}$$

Please note that now in (4) only $\boldsymbol{\Phi}$ depends on the choice of the pdf prior on the sources. The natural gradient update rule is $\boldsymbol{W}^{+}_{::k} = (\boldsymbol{I} + \mu\widetilde{\nabla}_{\boldsymbol{W}_{::k}} \boldsymbol{C}_{IVA})\boldsymbol{W}_{::k}$. The step size μ is fixed to 0.01.

3.2 Choice of the source prior

To end the derivation of this cost function, the joint pdf expressing the relationship between the frequency bins must be defined. We chose to use the zero mean multivariate Student-t distribution (5) since it has been shown to be well suited to speech signals [6].

$$\begin{aligned} P_{\boldsymbol{y}_n}(\boldsymbol{y}_{nt:};\boldsymbol{\Sigma}_n) &\propto (1 + \frac{1}{v}\boldsymbol{y}'_{nt:}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{y}_{nt:})^{-\frac{v+K}{2}}, \\ \frac{\partial \log P_{\boldsymbol{y}_n}(\boldsymbol{y}_{nt:};\boldsymbol{\Sigma}_n)}{\partial \boldsymbol{y}_{nt:}} &= -\frac{v+K}{2}\frac{\boldsymbol{\Sigma}_n^{-1}\boldsymbol{y}_{nt:}}{1+\frac{1}{v}\boldsymbol{y}'_{nt:}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{y}_{nt:}}. \end{aligned} \tag{5}$$

The covariance matrix $\boldsymbol{\Sigma}_n \in \mathbb{C}^{K\times K}$ is usually taken as the identity matrix, which implies that the time samples are assumed identically distributed. In (5), K is the number of frequency bins and v is the degrees of freedom for the Multivariate Student-t distribution. In this study, v is taken equal to T, the number of time samples in the D-STFTs.

3.3 Integration of activity information

In the AV-IVA algorithm, we make $\boldsymbol{\Sigma}^{AV}_{n,t}$ dependent on the activity of the n-th source, through the use of a visual voice activity detector described in the next paragraph. Therefore, time samples are no longer considered identically distributed. $\boldsymbol{\Sigma}^{AV}_{n,t}$ is defined as $\boldsymbol{\Sigma}^{AV}_{n,t} = \frac{1}{K}diag(\sigma^2_{nt},...,\sigma^2_{nt})$ with:

$$\sigma^2_{nt} = \begin{cases} 1, & \text{if the V-VAD of the } n-\text{th source is active at time } t \\ \epsilon, & \text{else} \end{cases}$$

The prior on the n-th source now models the non-stationarity of this source:

$$\frac{\partial \log P_{\boldsymbol{y}_n}(\boldsymbol{y}_{nt:}; \boldsymbol{\Sigma}_{n,t}^{AV})}{\partial \boldsymbol{y}_{nt:}} = -\frac{v+K}{2} \frac{(\boldsymbol{\Sigma}_{n,t}^{AV})^{-1} \boldsymbol{y}_{nt:}}{1 + \frac{1}{v} \boldsymbol{y}'_{nt:} (\boldsymbol{\Sigma}_{n,t}^{AV})^{-1} \boldsymbol{y}_{nt:}} \tag{6}$$

The matrix $\boldsymbol{\Sigma}_{n,t}^{O}$ is defined in a similar fashion, but the activity detector is ideal and given by an oracle. The matrix $\boldsymbol{\Sigma}_{n}^{IVA} = I$ is also defined for the reference IVA algorithm, in which all the sources are considered stationary (active at all time).

The variable $\epsilon \in]0, 1]$ allows to weight the contribution of the activity detector. Its value can be chosen accordingly to the accuracy of the activity detector. The more accurate the activity detector is, the closer to zero ϵ should be. Note that if $\epsilon = 1$, the AV-IVA is behaving exactly like the reference IVA. In the evaluation, ϵ was set to 0.05.

It must be pointed out that either for the IVA, the AV-IVA or the O-IVA, the proposed $\boldsymbol{\Sigma}_{n,t}$ definition assumes that all of the frequency bins carries the same power at a given time, since all the diagonal coefficients are equal. This is not a realistic model for a speech signal. To overcome this problem, each frequency bins of the observation is normalized before processing the separation. However, the total power carried by all observations in a frequency band is saved, and after separation, the total power carried by a frequency band across by all estimated source is restored to its original value. This is done because the mean power of a specific frequency band across all observations should not be modified by the separation step.

Visual voice activity detector: The audiovisual data that we used for this study are those that were used in [10]. These are composed of three time vectors. One is the speech signal itself, and width and height of the lips of the speakers during the locution. These visual features have been extracted from a video sampled at 50Hz. The V-VAD itself is based only on the evolution of the height of the lips. The data are first smoothed using a 80ms long mean filter and its derivative was computed. In a first step, the source activity is set to true when the absolute value of the derivative of the lips height is above a threshold. In a second step, all the detected silences shorter than 300ms are suppressed of the estimated activity to reduce the number of miss-detection.

The performances of this V-VAD can be evaluated by comparison with an oracle activity detector (O-VAD). In this case, this oracle activity detector is simply based on audio of the signal before mixing and is a thresholded version of the power profile of the audio signal. By taking the oracle as reference, we can define a false positive error rate (the V-VAD gives true and the O-VAD gives false) and a false negative error rate (the V-VAD gives false but the O-VAD gives true). In the results presented in section 4, the false negative rate is 3% and the false positive rate is 24%. Also, since the V-VAD inherently cannot detect silences shorter than 300ms, it may be more meaningful to not take into account these short silences. In this case, the false positive rate is 16%.

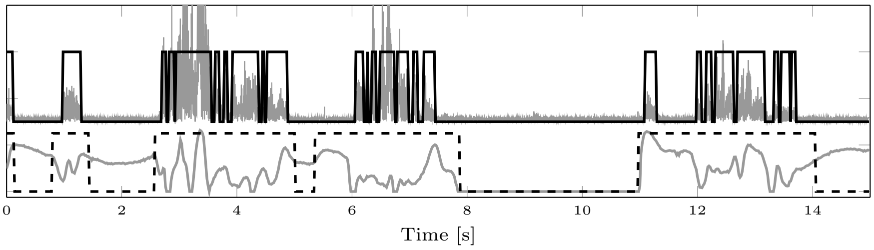


Fig. 1: Example of typical data used in this study. Top plot, Gray line : absolute value of speech signal. Black line : Oracle activity detector. Bottom plot : Gray line, height of lips. Black dashed line : Visual Voice Activity Detector

4 Experimental Results

The proposed method was evaluated over 39 trials. All of the figures presented below are averaged over these trials. At each trial, three different sources were randomly selected over a database of 77 fifteen seconds length speech samples. The mixing filters where generated at each trial from measured impulse responses found in the MARDY database [13]. These impulse responses where measured between 3 speakers and 24 different microphone positions and were truncated to 150ms in our study. Also, only the 8 microphones the farthest of the speakers were used. The mixtures were then generated by convolution in the time domain. The audio sampling rate was 8kHz and the STFTs were processed using 150ms frames with an overlap of 90%. The unmixing filters were initialized to identity (i.e. $\forall k \in [1, K]$, $\boldsymbol{W}_{::k} = I$).

The behaviour of the proposed method where evaluated over three different scenarios: a separation in a determined case where the activity of each source is estimated, an extraction in a determined case where only one source activity is estimated, and a extraction case where only one activity is estimated but in an underdetermined case where only two observation are available for three sources.

4.1 Performance Measure

The separation performance where evaluated with two criteria. The first one is the Signal to Interference Ratio (SIR) [12]. It measures the quality of extraction of a source. It is a signal to noise ratio where the signal is an estimated source, and the noise is the remaining signals of the other sources in the estimation of the first one. The higher the SIR is, the better the extraction is. In a simulation, the computation of the SIR is straight forward because we have access the separate contribution of each source to each estimated source after the separation stage. We denote $\tilde{\boldsymbol{s}}_{m,n}(t)$ the contribution of the n-th source to the m-th estimated source. The definition of SIR is then:

$$SIR_m = 10 \log \frac{\sum_t \tilde{\boldsymbol{s}}^2_{m,m}(t)}{\sum_t (\sum_{n \neq m} \tilde{\boldsymbol{s}}_{m,n}(t))^2} \tag{7}$$

The second criterion is the Identification Failure Rate (IFR). It measures how often each estimated source has been associated to the rightful activity after the separation stage. In the separation case it means that the output permutation should be the same than the input permutation of the sources. The permutation computed at the performance evaluation stage is the one maximizing the mean SIR of the estimated sources. In the extraction case, only the target source has to be rightfully aligned. The IFR is the failure rate on these criteria, so the lesser it is, the better is the identification performance.

4.2 Experiments

In this experiment, the proposed method, designated AV-IVA, is compared to a standard IVA algorithm, where no visual information is used. It is also compared to an oracle IVA designated O-IVA where the ideal VAD presented in section 3.3 is used. This allows to compute an upper limit on the performance of the proposed method. In the three methods, the same implementation of the IVA algorithm was used. Only the activity data carried by $\boldsymbol{\Sigma}_{n,t}$ was changed.

Determined separation case: In this experiment the goal is to recover all of the sources. So, the displayed SIR are the SIR averaged over the three estimated sources. To compute the IFR, we consider that the output order of sources is valid only if it is the same than the order of the source before the mixing step. Results in Fig.1.a and Table.1.a show that the proposed method improves both

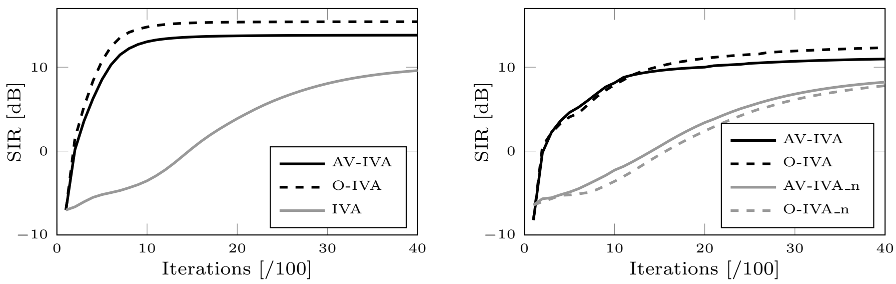
Table 1: SIR and IFR after convergence in the determined case

Prior	SIR start (dB)	SIR end (dB)	IFR
AV-IVA	-7.0	14	0%
O-IVA	-7.0	15	0%
IVA	-7.0	9.5	92%

(a) Separation case

Prior	SIR start (dB)	SIR end (dB)	IFR
AV-IVA	-8.2	11	3%
O-IVA	-8.2	12	3%
AV-IVA_n	-6.4	8.2	43%
O-IVA_n	-6.4	7.8	48%

(b) Extraction case



(c) Separation case.

(d) Extraction case.

Fig. 2: Determined case : Evolution of performance index during optimisation (SIR vs number of iterations)

the separation quality and the convergence speed. The permutation error is also dramatically reduced, but such a comparison might be unfair to the IVA. The output permutation of the IVA depends only on the mixing process because no other information is given before the separation step. Also as expected, the AV-IVA performances are below the O-IVA case.

Determined extraction case: In this experiment, the goal is to extract a specific speech from the determined mixture. Only the activity about this source is used. The other ones are considered active at all time (i.e. $\boldsymbol{\Sigma}_{n,t} = I \quad \forall n \neq 1$). To compute the IFR, we consider that only the target source should be aligned with its associated activity. Results in Fig.2.a and Table.2.a show that for the target source, SIR and IFR improvements are a bit inferior to the one in the separation case, and the extraction converges slower, but the results are still better than those of the reference IVA. The SIR and IFR of the non-target sources are behaving like the ones of the reference IVA in the separation case (The IFR are inferior to the ones in the separation case, but still correspond to the one given by a random permutation between sources).

Underdetermined Extraction case: In this section, the goal is to extract a specific speech from the mixture and only the activity information about this source is known. However, only two observations are used to perform this extraction while the mixture is composed of three sources. As shown in Fig.3 and Table.3, the target source is better enhanced than the other sources but the overall performances are much below the ones in the determined case. This is due to the fact that the mixing matrix are not invertible in an underdetermined mixture. While the target source can be enhanced, a complete extraction cannot be reached.

Table 2: Separation results in the under-determined case.

Prior	SIR start (dB)	SIR end (dB)	IFR
AV-IVA	-8.7	3.5	8%
O-IVA	-8.7	5.7	8%
IVA	-8.7	0.6	60%

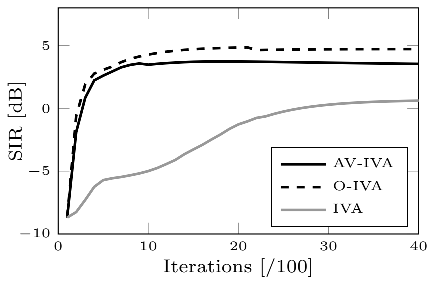


Fig. 3: Underdetermined case : Evolution of performance index during optimisation : (SIR vs number of iterations)

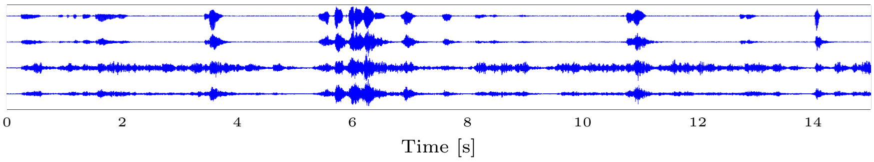


Fig. 4: Example of a speech signal enhancement using AV-IVA in an underdetermined mixture. From top to bottom : Original speech, Recording of speech alone in the reverberant environment, Mixture, Enhanced Speech.

An example of enhancement of a speech signal in an underdetermined mixture with real reverberation in presented in Fig.4. The mixture was composed of three sources recorded one after another by two microphones. The speaker were relocated in the room for each recording and the recordings were added afterwards to compose the mixture. For the target source, the SIR was -1dB at initialisation, and 2.5dB after enhancement.

4.3 Discussion

Several points need to be discussed about the proposed method. The first point is about the underdetermined case. While this technique cannot be used to achieve an exact extraction of the target speech in an underdetermined mixture, it still can be used as a preprocessing step to enhance the desired signal before using another extraction technique, like frequency masking (see [7]), or before a voice recognition software.

The second point is that the V-VAD proposed in this paper is based on segmented data of the face : the lips. These might be challenging to detect in a non-controlled environment. However, only the absolute value of the derivative of the mouth height is used. Using the norm of the optical flow computed below the eyes of the speaker should gives the same results, as long as the face of the speaker is detected. An accurate segmentation of the lips is therefore not needed.

A third point is that in this study, the activity information is used in the same way across all frequency bands because the activity detector gives only a binary information about the presence of the source at a certain time. However, this methods is based on making the covariance matrix of the frequency bins of a source dependent on time. This matrix can carry much more information about the frequency structure of the signal. If the available activity detector can give more information about the frequency structure of the desired source, it can also be used by the proposed method. An example would be the musical score of an instrument we want to enhance in a stereo recording containing several instruments. However, more studies shall be necessary to confirm that last point.

5 Conclusion

In this paper, we presented a way of using a visual voice activity detector to improve the separation or extraction of speech signals from a convolutive mixture using the IVA framework. A simple activity detector based on lip movements was presented and its output was included into a natural-gradient based IVA algorithm. Presented results show that the proposed method improves the separation of speech signals in a determined mixture, accelerate the convergence of the algorithm and allows to identify the target sources. The proposed method is also able to enhance a specific source in an underdetermined mixture.

References

1. Amari, S.i., Cichocki, A., Yang, H.H.: A new learning algorithm for blind signal separation. Advances in neural information processing systems pp. 757–763 (1996)
2. Anderson, M.: Independent Vector Analysis: Theory, algorithms, and applications. Ph.D. thesis (2013)
3. Cherry, E.C.: Some experiments on the recognition of speech, with one and with two ears. The Journal of the acoustical society of America 25(5), 975–979 (1953)
4. Hiroe, A.: Solution of permutation problem in frequency domain ICA, using multivariate probability density functions. In: International Conference on Independent Component Analysis and Signal Separation. pp. 601–608. Springer (2006)
5. Kim, T., Eltoft, T., Lee, T.W.: Independent Vector Analysis: An extension of ICA to multivariate components. In: International Conference on Independent Component Analysis and Signal Separation. pp. 165–172. Springer (2006)
6. Liang, Y., Chen, G., Naqvi, S., Chambers, J.: Independent Vector Analysis with multivariate student's t-distribution source prior for speech separation. Electronics Letters 49(16), 1035–1036 (2013)
7. Pedersen, M.S., Larsen, J., Kjems, U., Parra, L.C.: A survey of convolutive blind source separation methods. Multichannel Speech Processing Handbook pp. 1065–1084 (2007)
8. Rivet, B., Wang, W., Naqvi, S.M., Chambers, J.A.: Audiovisual speech source separation: An overview of key methodologies. IEEE Signal Processing Magazine 31(3), 125–134 (2014)
9. Sawada, H., Mukai, R., Araki, S., Makino, S.: 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)
10. Sodoyer, D., Rivet, B., Girin, L., Schwartz, J.L., Jutten, C.: An analysis of visual speech information applied to voice activity detection. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. IEEE (2006)
11. Summerfield, Q.: Use of visual information for phonetic perception. Phonetica 36(4-5), 314–331 (1979)
12. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE transactions on audio, speech, and language processing 14(4), 1462–1469 (2006)
13. Wen, J.Y., Gaubitch, N.D., Habets, E.A., Myatt, T., Naylor, P.A.: Evaluation of speech dereverberation algorithms using the MARDY database. In: in Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC). Citeseer (2006)