AN IMROVEMENT IN USING HERMITIAN ANGLEIN CONVOLUTIVE SPEECH BLIND SOURCE SEPARATION

Hamid Mahmoodian[1], Atefeh Soltani[2], Ali Hashemi[3]

[1]Electrical Faculty, Najafabad Branch, Islamic Azad University, Daneshgah Blvd., Najafabad, Iran
h_mahmoodian@pel.iaun.ac.ir
[2]Electrical Faculty, Najafabad Branch, Islamic Azad University, Daneshgah Blvd., Najafabad, Iran
atefe_soltani2004@yahoo.com
[3]Electrical Faculty, majlesi Branch, Islamic Azad University, Majlesi Town, Iran
a.hashemi@iaumajlesi.ac.ir

ABSTRACT

This paper presents a T-F masking method for convolutive blind source separation based on hermitian angle concept. The hermitian angle is calculated between T-F domain mixture vector and reference vector. Two different reference vectors are assumed for calculating two different hermitian angles, and then these angles are clustered with k-means or FCM method to estimate unmixing masks. The well- known permutation problem is solved based on k-means clustering of estimated masks which are partitioned to small groups. The experiment results show that separation performance for two different reference vectors is better than that for only one reference vector.

Index Terms— blind source separation (BSS), sparsity , w-disjoint orthogonality,hermitian angle.

1. INTRODUCTION

The blind source separation problem is extracting original signals from their mixtures, assuming there is not any information about mixing process or original signals. The mixing model is instantaneous or convolutive. The problem can be explained as follow:

Suppose that the source signals and microphone outputs are called as $s_1, s_2, \dots, s_q$ and $x_1, x_2, \dots, x_p$ convolutive BSS can be expressed as

$$x_p(n) = \sum_{q=1}^{Q} \sum_{l=0}^{L-1} h_{pq}(l) S_q(n-l) \quad (1)$$

where, P is the number of microphones, Q is the source number, $p = 1,..,P$ and $q = 1,..,Q$. L is the mixing filter length where the output signal vectors and p^{th} microphone output samples are shown as

$$x = [x_1, x_2, \dots, x_p]^T, x_p = [x_p(0), \dots, x_p(N-1)]^T$$

In the previous relation 'T' is transpose operator, N is the number of total samples column vector of sources and q^{th} source samples are defined as

$$s = [s_1, s_2, \dots, s_p]^T, \quad s_{q=}[s_q(0), \dots, s_q(N-1)]^T$$

The impulse response from q^{th} source to p^{th} microphone is $h_{pq}(l)$, $l = 0, \dots L$.Overdetermined and underdetermined BSS are related to the number of sources and sensors (microphones) comparing together. The separation criteria can be divided in to the methods based on higher order statistics (HOS) and second order statistics (SOS)[1]-[5]. In the underdetermined BSS (P< Q), SCA is the most popular method. There are several methods [6]-[9] that work based on the sparseness of the source signals. If the signals are sufficiently sparse, it could be assumed that the sources rarely exist simultaneously.

2. PROPOSED METHOD

2.1. Signal Transformation

In the first stage of proposed method in Fig. 1, time-domain mixture signals which are sampled at frequency f_s are transformed in to time-frequency domain using STFT analysis. Time-Frequency transformation of (1) is:

$$X(k,t) = H(k)S(k,t) \sum_{q=1}^{Q} H_q(k) S_q(k,t) \quad (2)$$

where $X(k,t)$ is T-F transformation of microphone output vectors and S(k,t) is the STFT of source signals .i.e.:

$$X(k,t) = [X_1(k,t), \dots, X_p(k,t)]^T$$

$$S(k,t) = [S_1(k,t), \dots, S_p(k,t)]^T.$$

The impulse response $H(k)$ and q^{th} source column vector of impulse response in the k^{th} frequency bin are:

$$H(k) = [H_1\ (k), \dots, H_q(k)]$$

$$H_q(k) = [H_{1q}\ (k), \dots, H_{pq}(k)]^T$$

where $H_{pq}(k)$ is the impulse response from q^{th} source to p^{th} microphone at the k^{th} frequency bin.

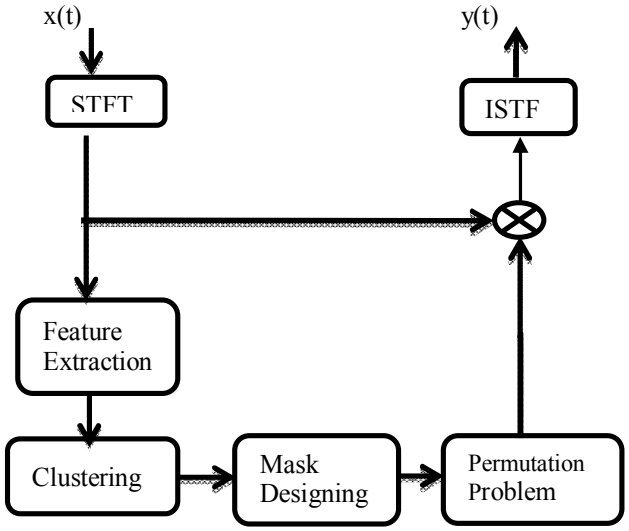

Fig.1 system overview

2.2. Feature Extraction

If the original sources were w-disjoint orthogonal, then the separation can be realized with gathering T-F points which belong one source. In our method, hermitian angle in complex vector space between STFT of microphone outputs and two reference vectors has been considered as the features in the procedure. In this method the feature space has been expanded in two dimensions comparing with proposed method in [10], which one vector has been considered as the reference. In complex space, the cosine of the complex-valued angle between two complex vectors is defined as [9]:

$$\cos(\theta_C) = \frac{u_1^H\, u_2}{\|u_1\|\|u_2\|} \quad , \ \|u\| = \sqrt{u^H\, u} \tag{3}$$

where u_1 and u_2 are two complex vectors and H is the complex conjugate transpose operation. $\cos(\theta_C)$ in (8) is defined as

$$\cos(\theta_C) = \rho e^{j\emptyset} \quad , \rho \le 1 \tag{4}$$

where

$$\rho = \cos(\theta_H) = |\cos(\theta_C)| \tag{5}$$

Using (9) and (10), two different angles are defined as

$$0 \le \theta_H \le \frac{\pi}{2} \quad , \quad -\pi \le \emptyset \le \pi$$

where θ_H is called hermitian angle and $\emptyset$ is named pseudo angle between two complex vectors u_1 and u_2. The hermitian angle between $H_q(k)$, impulse response vector and r, reference vector do not will change if it multiply by complex scalar (source signal $S_q(k,t)$) and it is same with hermitian angle between mixture vector, $X(k,t)$ and reference vector, where reference vector is a vector with P random element. So it could be a P element vector with all the elements equal to 1+j1. For P mixtures and Q sources, the hermitian angle between, each of the mixture vectors in the k^{th} frequency bin and reference vector with p element is explained as:

$$\cos(\theta_C(k1, t1)) = \frac{X(k_1,t_1)H_r}{\|X(k_1,t_1)\|\|H_r\|} \tag{6}$$

$$\theta_H^{(k_1)}(t_1) = \cos^{-1}(|\cos(\theta_C(k_1,t_1))|) \tag{7}$$

The original source signals $S_q, q = 1, \dots, Q$ are assumed sparse in T-F plane i.e. at any T-F point there is at most contribution of one source. The hermitian angle between mixture vector and reference vector is equal to the hermitian angle between impulse response $H_q(k)$ and r, reference vector, where $S_q(k,t)$ is corresponding to the present source at that point (k,t). The expanded space of features should be used to recognize the sources of the samples.

2.3. Clustering and Mask Design

By using Clustering algorithm to cluster the vector of hermitian angle in the k^{th} frequency bin, Q different clusters are produced such that, the samples which belong to one cluster are components of one source.

K-means and fuzzy c-means (FCM) have been used for clustering of hermitian angles. Using k-means method, the result mask produced with k-means algorithm, has binary values (0, 1) so it introduces artifact in reconstructed signals. FCM method is used for clustering the angles to produce soft and smoother masks with respect to the masks produced by k-means. After clustering Index vector of the k-means clustering ({0, 1}) and membership function of the fuzzy c-means clustering ([0,1]) have been used as masking values.

2.4. Permutation Problem

Blind source separation of signals in frequency domain has two important problems, the scaling and the permutation problem. Because of applying the generated T-F masks directly to the mixtures, scaling problem does not exist in our experiment. To solve the permutation problem, many algorithms have been reported in the frequency domain [11],[12].

For speech signals in T-F domain, increasing the distance between two frequency bins, decreasing the correlation between them. Therefore, frequency bins are portioned to small groups that signal for one frequency group is more correlated which in turn; the permutation problem is solved for each small group [10]. The groups are composed

of few adjacent frequency bins with overlap which in our experiment, 16 bins with 75% overlap are assumed. To solve the permutation problem for the groups, the masks are clustered by k-means method to minimize the following equation:

$$D = \sum_{q=1}^{Q} \sum_{\substack{M_i^k \in C_q \\ i=1,\dots,Q \\ k=k_{st},\dots,k_{end}}} \left(1 - r_{M_i^{(k)} C_q}\right) \quad (8)$$

where M_i^k is the i^{th} mask in the k^{th} bin frequency and C_q is the center of q^{th} cluster, and $r_{M_i^{(k)} C_q}$ is the correlation between them. The parameters, k_{st} and k_{end} are starting and ending frequency bins for each frequency group.

The k-means categorizes the masks in each group into Q clusters based on distance metric correlation to have high correlated masks in each. Ideally, for any frequency bin, each cluster has exactly one mask, but it does not occur in practical situation. For instance, at some frequency bins, some clusters do not have any mask and some clusters have more than one mask. In such bins the k-means is failed to cluster them correctly. In such case, correlation between the centers and the masks has been calculated to maximize the summation of correlation values. The permutation matrix π_k in the k^{th} bin for failed clusters is defined as

$$\pi_k = \arg_\pi \max \sum_i^F \sum_j^F (\pi . R_{CM})_{ij} \quad (9)$$

where '$\bullet$' is the element wise product between two matrixes, F is the number of failed clusters, π is a selected permutation matrix with elements 0 and 1, such that there is one '1' in each row or column. In this case permutation matrix could be selected as $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ or $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. $R_{CM} \in R^{F \times F}$ is correlation matrix, $(R_{CM})_{ij}$ is the Pearson correlation between i^{th} row of C and j^{th} row of M and q is a number which belongs to the index of failed clusters. C and M are respectively the center matrix and the mask matrix in the failed clusters in the k^{th} frequency bin, which are defined as

$$C = \left[\dots, C_q^T, \dots\right]^T \ , C_q \in R^F$$

$$M = [\dots, M_q^T, \dots]^T \ , M_q \in R^F$$

and therefore the masks matrix after solving permutation in the k^{th} frequency bin will be $\pi_k M$.

2.5. Reconstruction of Output Signals

After Time-Frequency mask generation, the separated signals are produced by

$$Y_q(k,t) = M_q(k,t) X_P(k,t), \forall t, \quad q = 1, \dots, Q \quad (10)$$

where $X_P(k,t)$ is one of the mixtures ($p = 1$ or 2). Finally, inverse STFT (overlap and add method reported in [13]) has been used to produce the separated signals and come back it to time domain. The generated signal is the estimation of original signals.

3. EXPERIMENTAL RESULTS

3.1. Dataset

For performance evaluation of the proposed method some criterions are needed. Room simulator software version 1 is used to simulate a room conditions and control the microphones and sources positions. Five different Signals from TIMIT database have been used to mix in the simulator and make the convolutive speech signals.

3.2. Performance Criterions

The separation quality is measured by method suggested in [14], [15]. In this method, the estimated signals are written as

$$y_q = y_{qtarget} + e_{interf} + e_{qartif} \quad (11)$$

where $y_{qtarget}$ is target signal with allowed deformation, e_{interf} is interference due to unwanted sources and e_{qartif} is artifacts introduced by the separation algorithm. So the performance criterions are defined as

$$SDR = 10 \log_{10} \frac{\|y_{qtarget}\|^2}{\|e_{qinterf} + e_{qartif}\|^2} \quad (12)$$

$$SIR = 10 \log_{10} \frac{\|y_{qtarget}\|^2}{\|e_{qinterf}\|^2} \quad (13)$$

$$SAR = 10 \log_{10} \frac{\|y_{qtarget} + e_{qinterf}\|^2}{\|e_{qartif}\|^2} \quad (14)$$

SDR is source to distortion ratio, SAR is source to artifact ratio and SIR is source to interference ratio. Since the masks are applied to p^{th} microphone output, the target signal is $y_{qtarget} = h_{pq} * s_q$, where h_{pq} is the impulse response from p^{th} source to q^{th} microphone.

3.3. Experiment Results with One Reference Vector

The reference vector is two dimensional vector with the elements 1+j, and the separation performances (SP) in (db) of the simulation for two clustering methods have been shown in table I and II.

3.4. Experiment Results with Two Reference Vector

Based on our proposed method, the hermitian angle is calculated with respect to two different reference vectors. The elements of the first and second references vector have been selected as 1+j and a random value respectively. The

[1].[Online] . Available: http://bass-db.gforge.inria.fr/BASS-dB/?show=browse&id=filters

separation performance (SP) for two reference vectors are shown in table III and IV. The comparison between one and two references vector shown in tables, illustrate that the latter has improvement in both SDR and SAR by FCM clustering method although k-means has more improvement in SIR. Comparing the results shows that the signal separation with 20cm microphone distance is better than one with14cm

4. CONCLUSION

T-F masking method has been used for convolutive blind source separation of speech signals such as previous work [10]. In this work, hermitian angles with respect to two reference vectors has been proposed and results show the superiority of the SDR, SIR and SAR comparing with one reference vector has been used.

TABLE I
SEPARATION PERFORMANCE FOR ONE REFERENCE VECTORMICROPHONE DISTANCE=14CM

SP	Input	k-means		FCM	
		Output	Improv.	Output	Improv.
SDR	0.06	6.2	6.14	6.51	**6.45**
SIR	0.06	10.64	**10.58**	9.6	9.54
SAR	20	8.83	-11.17	10	**-10**

TABLE II
SEPARATION PERFORMANCE FOR ONE REFERENCE VECTORMICROPHONE DISTANCE=20CM

SP	Input	k-means		FCM	
		Output	Improv.	Output	Improv.
SDR	0.08	6.36	6.28	7	**6.92**
SIR	0.08	10.6	**10.52**	10.26	10.18
SAR	20	8.9	-11.1	10.4	**-9.6**

TABLE III
SEPARATION PERFORMANCE FOR TWO REFERENCE VECTORMICROPHONE DISTANCE=14CM

SP	Input	k-means		FCM	
		Output	Improv.	Output	Improv.
SDR	0.06	9.2	9.14	10.1	**10.04**
SIR	0.06	14	**13.94**	13.56	13.50
SAR	20	11.8	-8.2	13.3	**-6.7**

TABLE IV
SEPARATION PERFORMANCE FOR TWO REFERENCE VECTORMICROPHONE DISTANCE=20CM

SP	Input	k-means		FCM	
		Output	Improv.	Output	Improv.
SDR	0.08	10.79	10.71	10.98	**10.9**
SIR	0.08	15	**14.92**	14.3	14.22
SAR	20	13.43	-6.57	13.87	**-6.13**

5. REFERENCES

[1] V. Capdevielle, C. Servire, and J. L. Lacoume,"Blind separation of wide-band sources in the frequencydomain," in *ICASSP95*, vol. III, Detroit, May 9–12, pp. 2080–2083, 1995.

[2] S. Icart and R. Gautier, "Blind separation of convolutive mixtures using second and fourth order moments," in *ICASSP'96*, vol. 5, pp. 3018–3021, 1996.

[3] D. Yellin and E. Weinstein, "Criteria for multichannel signal separation," *IEEE Trans. Sig. Proc.*,vol. 42, no. 8, pp. 2158–2168, Aug 1994.

[4] P. Comon, E. Moreau, L. Rota, "Blind separation of convolutive mixtures: A contrast-based joint diagonalization approach," in *ICA'01*, pp. 686–691, 2001.

[5] J. Thomas, Y. Deville, and S. Hosseini, "Timedomain fast fixed-point algorithms for convolutive ICA," *IEEE Sig. Proc. Lett.*, vol. 13, no. 4, pp. 228–231, Apr 2006.

[6] O. Yilmaz and S. Rickard, "Blind separation ofspeech mixtures via time-frequency masking," *IEEETrans. Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, Jul2004.

[7] N. Roman, "Auditory-based algorithms for soundsegregatoion in multisource and reverberant environtments,"Ph.D. dissertation, The Ohio State University,Columbus, OH, 2005.

[8] A. Blin, S.Araki, and S.Makino, "Underdeterminedblind separation og convolutive mixtures of speechusing time-frequency mask and mixing matrix estimation,"*IEICE Trans. Fundamentals*, vol. E88-A,no. 7, pp. 1693–1700, Jul 2005.

[9] S. Xie, L.Yang, J.Yang, G.Zhou, and Y. Xiang ,"Time frequency Approach To Underdetermined Blind Source Separation,"IEEETransaction onneuralnetworksandlerningsystems, vol. 23, no. 2, february 2012.

[10] V. G. Reju, S. N. Koh and I. Y. Soon, "Underdetermined Convolutive Blind SourceSeparation via Time-Frequency Masking," IEEE Transactions on Audio, Speechand Language Processing, Vol. 18, NO. 1, pp. 101–116, Jan. 2010.

[11] S. Dubnov, J. Tabrikain, and M. Arnon-Targan, "Amethod for directionally-disjoint source separationin convolutive environment," in ICASSP'04, vol. V, pp. 489–492, 2004.

[12] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in ICA'06, pp. 601–608, 2006.

[13] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, Discrete-Time Signal Processing. Saddle River, NJ: Prentice-Hall, 2003.

[14] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[15] C. Fevotte, R. Gribonval, and E. Vincent, "BSS EVAL Toolbox User Guide, IRISA Tech. Report 1706", Tech. Rep. Rennes, France, 2005 [Online]. Available: http://www.irisa.fr/metiss/bss eval/