BLIND SEPARATION OF MOVING SPEECH SOURCES USING SHORT-TIME LOD BASED ICA METHOD

Jing Zhang, P.C. Ching

Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong
jzhang@ee.cuhk.edu.hk , pcching@ee.cuhk.edu.hk

ABSTRACT

This paper describes the application of an effective short-time ICA method for blind speech separation of a moving-speaker system. For the situation where time-varying mixture exists, adaptive ICA techniques encounter difficulties due to the collapse of sources independence assumption under short-time analysis. In this paper, we propose a method based on the short-time local optima distribution (LOD) of feasible separation region to alleviate such problems. Based on the characteristics of these distributions, information is obtained for avoiding local traps and approaching the desired global optimum of the de-mixing matrix. Simulation tests of the proposed method show its effectiveness for blind separation of moving speech sources.

Index Terms—blind speech separation, speaker moving, short-time LOD, ICA

1. INTRODUCTION

The state of the art blind speech separation technology is still vulnerable in real acoustic environment. Especially one of the most difficult problems encountered is from competing moving speakers, or even worse, the target speaker's location is also varying with time. The mixing system becomes time varying for such scenarios.

Independent component analysis (ICA) is often applied for blind speech separation to recover sources by making the separated outputs as independent as possible, with the lone assumption that the sources are mutually independent. A variety of successful ICA methods have been developed for this purpose [1], but only very little has been studied regarding source-moving situations [2-4]. For batch algorithm, the time-varying mixing matrix is assumed to be constant in a short time interval [3], which leads to an important but often overlooked drawback that the sources independence assumption collapses for such short time analysis. It will often cause the global optimum obtained during search to be different from the desired point, and sometimes even stuck in local traps, as what commonly happens during gradient-based training [5]. These two problems greatly degrade the performance of adaptive ICA techniques for separating moving sources.

In this paper, we first address the problem related to blind speech separation of moving speakers. Three basic local optima distribution types are represented with respective characteristics. Based on these characteristics, the short-time LOD based ICA method is proposed and its effectiveness in achieving a faster and more accurate short-time analysis will be demonstrated.

2. THE PROBLEM IN SOURCE-MOVING SEPARATION TASK

Consider an instantaneous mixture of a two-speaker-two-sensor system $X = AS$, where $S=[s_1 \;\; s_2]^T$ represents two speech sources, $X=[x_1 \;\; x_2]^T$ represents two mixed observations, and $A=[a_{11} \;\; a_{12};\; a_{21} \;\; a_{22}]$ is the mixing matrix.

ICA works for estimating the desired de-mixing matrix $W=[w_{11} \;\; w_{12};\; w_{21} \;\; w_{22}]$, such that the separated outputs $Y = WX$ ($Y=[y_1 \;\; y_2]^T$ represent two separated signals) are as independent to each other as possible.

The de-mixing matrix W is time varying with source moving. For traditional batch algorithm, the mixed signals are divided into N short-time intervals. A time scale T is chosen as the length of such intervals, in which a good tradeoff between the following two issues is attempted: 1) the mixture can be assumed fairly stationary; and 2) the sources independence assumption holds. Then the separation is carried out on every short-time interval $[t, t+T-1]$ to obtain $[W_1, W_2 \cdots, W_N]$ for approximating the actual time-varying de-mixing matrix W .

It can be seen that the separation performance for moving sources depends highly on this length T [3]. To make the estimated mixture close to the time-varying mixture for accurate separation, T is preferred to be as small as possible, but problem might occur because the sources are no longer mutually independent, and the floating global optimum might end up in local sub-optimal traps.

We have studied the blind speech separation in short-time interval with quite small length (like 20ms with 320 samples), of which very little study could be found in current literature. Instead of just compromising the two issues, the proposed method works for the blind separation of moving speech sources with more accurate mixture approximation, and is capable to alleviate the problems when sources independence collapses. Details will be given in the following paragraphs.

3. THE PROPOSED METHOD

3.1 Pre-processing

The de-mixing matrix parameters are used as the decision vector for passive covering in building LOD. Considering the difficulties in analyzing LOD with high dimensions, we first perform the *decision vector dimension reduction* based on the permutation and scaling ambiguities of ICA [6]. Since the mixing matrix A is non-singular, and the order of sources is not important, we can set $|a_{11}|>|a_{12}|, |a_{22}|>|a_{21}|$. Let:

$$\begin{cases} s_1' = a_{11}s_1 \\ s_2' = a_{22}s_2 \end{cases} \quad (1)$$

$$\alpha = \frac{a_{12}}{a_{22}}, \beta = \frac{a_{21}}{a_{11}}, |\alpha|<1, |\beta|<1 \quad (2)$$

Thus, the mixing signals can be rewritten as follows:

$$\begin{cases} x_1 = a_{11}s_1 + a_{12}s_2 = s_1' + \alpha s_2' \\ x_2 = a_{21}s_1 + a_{22}s_2 = \beta s_1' + s_2' \end{cases} \quad (3)$$

Therefore, the de-mixing matrix can be transformed into $W = [1 \quad p; q \quad 1]$ with desired $p = -\alpha, q = -\beta$. Thus, the decision vector is $[p, q]$ ($p, q \in (-1,0) \cup (0,1)$).

3.2 LOD

The K-L divergence [7]: $D_{KL} = \int_{-\infty}^{+\infty} f(y) \log(\frac{f(y)}{q(y)}) dy$ is used here as a measure of independence, where $f(y)$ represents the joint probability density function (PDF) of the separated signals, and $q(y)$ represents the product of the separated signals' marginal PDF. D_{KL} becomes smaller for higher independence.

We apply passive covering [9] to get the contour plot of feasible separated signals' independence measure with the decision vector $[p, q]$. It is called the "local optima distribution" (LOD), which represents the distribution shape of the feasible local optima (shown by black points) with D_{KL} decreasing along the direction of arrow, as seen in Fig.1 and Fig.2(a).

Experiments have been carried out on 13344 sets of short-time speech segments extracted from the TIMIT database (4 sets of speakers $\times$ 24 sets of mixtures $\times$ 139 short-time frames of 320 samples). Results showed that: with limited data, the global optimum tends to drift away from the desired optimum, and local optima with low D_{KL} appear and diffuse from the quadrant with desired point (*desired quadrant*) to adjacent ones. Different local optima distribution types emerge and bring about difficulty for separation.

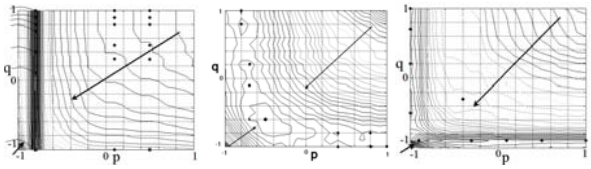
Notice that the comparative energy and kurtosis of sources, as well as the mixing matrix affect the LOD, which is easily understood by the relationship of second- and fourth-order cumulants and the signal's PDF that determines the D_{KL} value. Based on such effects, three basic LOD types are proposed for roughly covering various cases in instantaneous mixture.

LOD type 1:

Dominant distribution (DoD) as illustrated in Fig. 1 (a), (c); If the effect of one source (*dominant source*) to the LOD shape is much higher than that of the other, the local minima with low D_{KL} appear approximately around one special value of p (or q), which directly decides the *dominant source* and therefore is called the "*dominant parameter"*. It is called *p-dominant (or q-dominant)* distribution. The global optimum is often far away from the desired one. Therefore the sequential algorithm is preferred [8] because we can achieve the *dominant source* right away.

LOD type 2:

Best opposite worse distribution (BOWD) as illustrated in Fig.1(b); If the effects of the two sources to the LOD shape are similar, the global optimum is often around the desired point, and the local optima with low D_{KL} emerge not only in the *desired quadrant* but also in the two adjacent quadrants, each having possible points for getting one of the sources. This is called the BOWD type in which a simultaneous algorithm is preferred [8]. Notice that here the quadrant with highest average independence measure (*worst quadrant*) is just opposite to the *desired quadrant*. It helps us to shrink the original feasible search region (FSR) to just the *desired quadrant* and therefore avoid the confusing local optima in the two adjacent quadrants. The accuracy and speed for search are improved consequently.



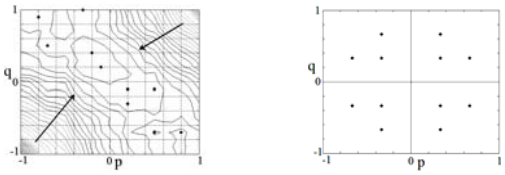
(a) p-dominant (b) BOWD (c) q-dominant

Figure 1: LOD of type 1 and type 2

LOD type 3:

Central Habitat distribution (CHD) as illustrated in Fig.2(a); If the absolute value of at least one of p and q is quite close to 0, the global optimum may embed in all local ones around the central part of the feasible search region. This case is actually the extension of the BOWD type but affected by the desired optimum with quite small absolute value, which drags the region with low D_{KL} to the central part. Therefore the possible initial points (PIP) may be chosen as shown in Fig.2(b). It is quite helpful for the

"blind search" task, in which the desired optimum is not known a priori and yet it is critical to avoid possible local traps. The simultaneous algorithm is still preferred [8] with the projected PIP as starting points.



(a) CHD (b) PIP

Figure 2: LOD of type 3 and PIP for such type

3.3 The LOD Based ICA Method

Fig.3 shows the proposed method based on the characteristics of different LOD types analyzed as above. LOD affords extra information and therefore helps to improve the performance of blind speech separation with speaker moving by short-time ICA analysis. Details of the deflation algorithm are presented in [8], and the Infomax method [1] is used as the simultaneous ICA for BOWD and CHD types.

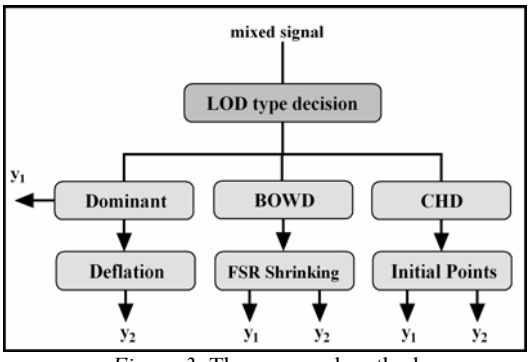


Figure 3: The proposed method

4. SIMULATION TESTS

The simulation tests are carried out to demonstrate that the proposed method can separate moving speech sources in short-time interval with small length T, which means that fairly limited sample points are required to better approximate the source moving environment. In the meantime the collapse of sources independence is tackled by the proposed LOD-based ICA method.

The system is an instantaneous mixture of one male and one female speaker from the TIMIT database. 7680 data is used with sampling frequency of 16 kHz and equal signal power. The mixing matrix is $A=[1\ a;b\ 1]$ with a,b changes gradually from 0.1 to 0.8. As such, it is expected that a moving-time-frame implementation is needed to approximate the mixture changes and perform the actual separation. This is empirically designed after testing all the three LOD types. Based on part 3.2, the LOD type is determined by the position difference D_{pos} of all local optima with low D_{KL} in p or q of the feasible region. For the simulation tests, if $D_{pos} \le 0.1$, it is set to be the dominant type; if $0.1 < D_{pos} \le 0.2$, it is type 2; otherwise it will be included as type 3.

We first carried out the separation by the original ICA using Infomax method [1] for contrasting as illustrated in Fig.4 and the results by the proposed method in Fig.5, with 320, 640, 960 and 1280 as the frame length and 160 as the frame shift. The separation performance of each frame is represented by:

$$SIR(y_i) = 10\log\frac{\sum_t |s_i'(t)|^2}{\sum_t |y_i(t) - s_i'(t)|^2}(dB), i = 1,2 \qquad (4)$$

(t :data number in a frame. SIR larger than 30dB is set to be 30 dB for easier observation of other smaller SIR). For Fig.4, SIR smaller than -10dB is cut off the plot, and for fig.5, SIR smaller than 0dB is cut off the plot.

In Fig.4, for longer frame length, the separation performance degrades because of the worse time-varying mixture approximation. Smaller frame length has the advantage of more invariable mixture, but exposing to the problem of sources dependence. Thus the floating of global optimum and existence of local optima traps give quite varying performance during the search process.

Contrasting with the separation performance by the proposed method shown in Fig.5, LOD affords extra information for better approaching the desired point and avoiding terrible local optima. Therefore the smaller frame length of 320 and 640 can give better separation performance than longer frame length. The disadvantage from sources independence collapse with shorter frame length is therefore overcome.

The desired de-mixing matrix parameter p and q should gradually changes from -0.1 to -0.8. The estimated p and q values are shown in Fig.6. For p-dominant, only p value is given and q is set to 0 (similarly for q-dominant, only q value is given and p is set to 0) because the deflation process is applied for dominant LOD types.

From the experimental results, since we applied the proposed method by using the information from LOD, good separation performance of every frame (short-time interval) is achieved, therefore reveals the effectiveness of the proposed method for short time analysis with more accurate speech separation of moving speech sources.

5. DISCUSSION & CONCLUSION

We proposed an effective short-time ICA method based on the local optima distribution with limited number of data samples. This works for the blind speech separation with

speaker moving by batch algorithm, especially for the rapid moving cases (limited samples can be used for every short-time interval), or cases when the mixture estimated in the foregoing interval(s) offers little knowledge for the latter interval's mixture estimation.

Based on the LOD of the feasible separation region, the proposed method provides information for adaptive ICA search on (1) simultaneous or sequential algorithm choosing, (2) FSR shrinking, and (3) possible initial point selecting. In this way, the undesirable effects caused by sources independence collapse are alleviated. Comparing with the current methods for source-moving problem, our approach performs much better on short-time interval with smaller length T (which means much more limited sample points for use), therefore capable to achieve more accurate approximation of mixture changes due to real-time speaker moving. Although only limited data is available, good separation performance can still be obtained, because we apply the information extracted not only from the optimal points, but also from the distribution of them, which is helpful for avoiding local traps and approaching the desired global optimum of the de-mixing matrix. Future study of this method will be focused on extension to higher-dimensional situations.

6. REFERENCES

[1] Radu Mutihac, Marc M. Van Hulle, "a comparative survey on adaptive neural network algorithms for independent component analysis", Romanian Reports in. Physics, Vol. 55, Nos. 1-2, 2003

[2] Ryo Mukai, Hiroshi Sawada, "blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction", IEICE trans. Fundamentals, vol. E87-A, No.8, August 2004

[3] A. Koutras, "blind speech separation of moving speakers using hybrid neural networks", Eurospeech 2001

[4] Kenneth E. Hild II, etc, "blind source separation of time-varying instantaneous mixtures using an on-line algorithm", Proc. IJCNN 2001, pp. 424-429,2001

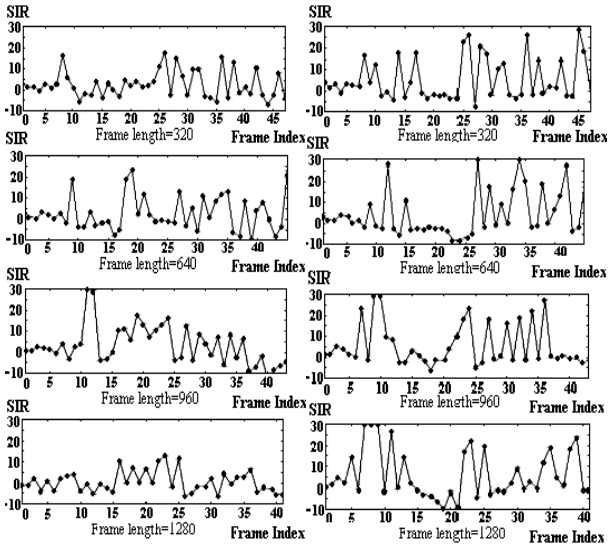
[5] Dinh-Tuan Pham, F.Vrins, "local minima of information-theoretic criteria in blind source separation", IEEE, signal processing letters, vol.12, No. 11, November 2005

[6] Ella B. and Aapo H., "a fast fixed-point algorithm for independent component analysis of complex valued signals", International Journal of Neural Systems, Vol. 10, No.1, (February, 2000) 1-8

[7] C.H. Chen, "on information and distance measures, error bounds and feature selection", Information Science, Vol. 10, pp. 159-173, 1976.

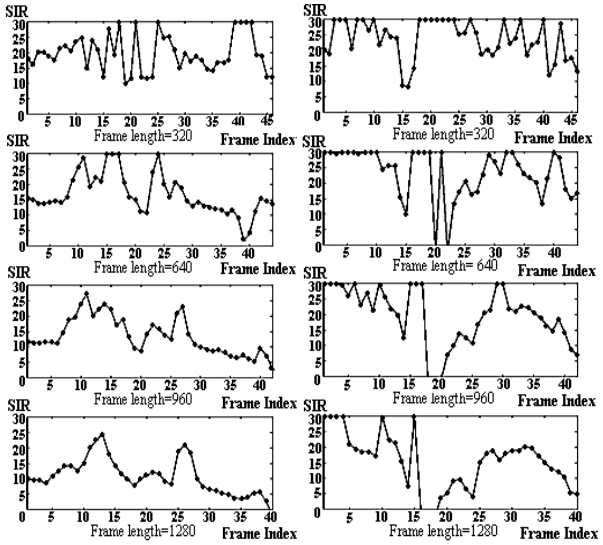
[8] A. Cichocki, S. Amari, "adaptive Blind Signal and Image Processing: learning algorithms and applications", pp: 178-179, 182, 191-193. New York: J. Wiley, c2002.

[9] D.P.Solomatine, "Genetic and other global optimization algorithms – comparison and use in calibration problems", Proc. 3rd Intern. Conference on Hydroinformatics, Copenhagen, Denmark, 1998. pp: 1021-1028



(a) SIR (y_1) of every frame (b) SIR (y_2) of every frame

Figure 4: SIR of separated outputs in frame by original ICA algorithm of Infomax



(a) SIR (y_1) of every frame (b) SIR (y_2) of every frame

Figure 5: SIR of separated outputs in frame by the proposed method

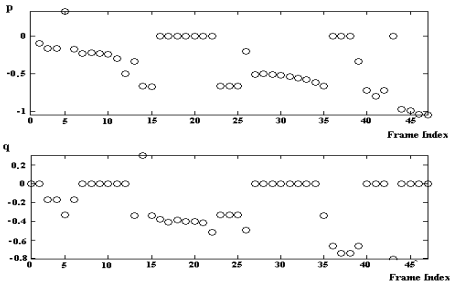


Figure 6: The estimated p values (upper) and the estimated q values (lower)