Short-Time ICA for Blind Separation of Noisy Speech

Jing Zhang, P.C. Ching
Department of Electronic Engineering
The Chinese University of Hong Kong, Hong Kong
jzhang@ee.cuhk.edu.hk, pching@ee.cuhk.edu.hk

Abstract. Noisy ICA model offers a viable solution for blind speech signal separation under real life scenarios, but it also creates new problem which need to be tackled, namely: estimating of the de-mixing matrix with non-invertible model. In this paper, a new algorithm based on short-time ICA is proposed for accurate estimation of the de-mixing matrix in the presence of noise. Without any assumption or prior knowledge of the noise covariance, the derivation of the proposed algorithm solely depends on the fact that the distribution shape remains pretty much unchanged although the estimated optima of the de-mixing matrix drift with noise contamination. In this work, we have set up a criterion on the distribution decision and signal-frame-selection for more accurate estimation. Experiment on Mandarin noisy speech showed quite satisfactory performance of matrix-estimation, thereby allowing better blind speech signal separation under noisy environment.

Keywords: Blind speech signal separation, noisy ICA model, short-time local optima distribution, signal-frame-selection

1 Introduction

For speech signal analysis, speech signal separation is necessary to solve the cocktail party problem [1] for applications like real-time automatic speech recognition, hearing aid, etc. Independent Component Analysis (ICA) method is often used for this goal. ICA aims at recovering individual 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 [2]. Although a noise-free model is usually applied for analysis, linear ICA model with additive Gaussian noise is frequently considered in many practical cases. By considering the noise added to the sources and/or sensors, the noisy-model approaches the reality much better than the noise-free alternate.

Not many methods for tackling noisy ICA exist in literature [3]. In practice, pre-processing (like simple filtering of time-signals, and PCA, etc) can be used to reduce noise in this respect. Literatures on the study of noisy model often assume that: (1) the contaminated noise is independent from the source signals and (2) the noise is Gaussian distributed. Higher-order cumulants are usually applied because they are immune to Gaussian noise[4,5]. However higher-order cumulants are sensitive to outliers, therefore the cumulants of orders higher than four are unlikely to be very
useful in practice. There is also the bias removal technique [6], which modifies the noise-free ICA method so that the bias due to noise is removed or at least reduced. It includes the de-mixing matrix estimation for achieving the separated noisy independent sources, and then the post-processing of de-noising them. Bias removal is a promising method but with an additional assumption that the noise covariance matrix is already known.

In this paper, we propose a new algorithm of the noisy independent component analysis model for blind speech signal separation. It is based on the fact that for short-time ICA, although the optima of the de-mixing matrix do float around with noise, the distribution of them is more robust against contamination. Especially for the so called “dominant local optima distribution”, which tends to hold its shape even with the presence of noise. Based on this characteristic, we can achieve good estimation of the de-mixing matrix for the noisy model, and therefore obtain better separated speech sources with noise for further de-noising process. No prior knowledge of the noise covariance is needed in this case.

In section 2, the definition of our noisy model for analysis is introduced. Then in section 3 we will describe the ‘dominant local optima distribution’ in noise-free model, and the factors that help to build up such kind of distribution, we shall also examine how such distribution retains its shape in noisy environment. A criterion for deciding such distribution for signal-frame-selection is given in section 4. The proposed algorithm is based on the analysis of the dominant local optima distribution and the decision criterion. Characteristics of Mandarin that help to establish the dominant local optima distribution to be built are discussed, and also experimental results are given to illustrate the effectiveness of the algorithm in noisy ICA model de-mixing matrix estimation for blind separation of noisy speech.

2. Definition of Noisy Model

We consider the noisy independent component analysis model for blind speech signal separation:

\[ x = As + n \] (1)

where \( s \) represents the speech source, \( x \) represents the mixing signal, \( A \) is an unknown mixing matrix, and \( n \) is some additive noise. We can rewrite the equation (1) into the following form:

\[ x = A(s + \tilde{n}) = \tilde{A}s \] (2)

where \( \tilde{n} = A^{-1}n \), assuming that the mixing matrix \( A \) is invertible, which is a basic assumption that leads to an ordinary ICA model. Although the independent components are noisy here, we first try to achieve a good estimation of the mixing matrix \( \hat{A} \), and from there, we will derive the scaled and/or permuted version of the
noisy speech sources $s + n$ by $A$. The estimation of the original speech from the separated noisy speech is an additional problem. Several methods for speech signal enhancement (e.g. spectral subtraction) can be used for this purpose.

3. Dominant Local Optima Distribution

Consider the linear instantaneous mixture of a two-source-two-sensor system by model (1). Now $s = [s_1 \ s_2]^T$, $x = [x_1 \ x_2]^T$, and $A = [a_{11} \ a_{12}; a_{21} \ a_{22}]$. Instead of directly estimating $A$, we estimate the desired de-mixing matrix $W = [w_{11} \ w_{12}; w_{21} \ w_{22}]$, such that the separated outputs $y = Wx$ ($y = [y_1 \ y_2]^T$) are as close to the sources (with noise in this case) as possible. It is easy to see that if $W = A^{-1}$, then $y = s + n$. In practice, permutation and scaling effects are present for all ICA [7], which, however, will not affect the speech separation goal.

3.1 Pre-processing

The de-mixing parameters that affect the separated outputs are used as the decision vector for passive covering in our work. Considering the difficulties in analyzing the local optimal distributions with high dimension, we first perform the decision vector dimension reduction based on the permutation and scaling ambiguities of ICA. 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:

$$
\begin{align*}
    s_1 &= a_{11}s_1 \\
    s_2 &= a_{22}s_2
\end{align*}
$$

(3)

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

(4)

Thus, the mixing signals can be expressed as:

$$
\begin{align*}
    x_1 &= a_{11}s_1 + a_{12}s_2 = s_1' + \alpha s_2' \\
    x_2 &= a_{21}s_1 + a_{22}s_2 = \beta s_1' + s_2'
\end{align*}
$$

(5)

The separation is unaffected since $s_1', s_2'$ are the scaled forms of $s_1, s_2$ respectively. Therefore, the de-mixing matrix can be transformed into $W = [1 \ p; q \ 1]$ with
desired parameters: \( p = -\alpha \), \( q = -\beta \) whilst the decision vector \([p, q]\), \( p, q \in (-1, 0) \cup (0,1) \).

3.2 Local Optima Distribution and the Dominant LOD Characteristics

![Fig1: LOD](image)

We apply passive covering to obtain the contour plot of the separated signals’ independence versus the decision vector \([p, q]\). The independence measure used is the K-L divergence \(D_{KL} \) [3] (the \(D_{KL}\) values decrease along the direction of the arrows). We call it local optima distribution (LOD) to express the distribution shape of local optima in the contour plot.

In order to look for the local optima distribution rules, experiments have been carried out on 4968 sets of short-time speech signals from the TIMIT database (3 sets of speaker sources \( \times \) 24 sets of mixing matrix \( \times \) 69 short-time frames with 320 samples). We found that the comparative energy and kurtosis of sources, as well as the mixing matrix parameters can affect the distribution shape of the local optima. It is easily understood by the relationship of second- and fourth-order cumulants and the signal’s PDF for independence measurement. Experimental results showed that: if one source has much lower energy and/or much higher kurtosis, then the so-called dominant distribution (shown in fig. 1(a) and (b)) tends to occur with the parameter that decides this source to be the dominant parameter. Therefore the increasing difference of the source’s absolute energy and kurtosis helps to build up such dominant distributions.

Detailed definition and decision criterion for dominant distributions will be given in section 4. But from fig.1, we can notice that the dominant distributions have
distinct characteristic in which one of the desired de-mixing matrix parameters, namely the dominant parameter, can be obtained right away. Also in the dominant distribution, the dominance of the source in energy and kurtosis is so obvious that even when noise exists, the distribution shape can be retained fully. While for other LODs, noise effect is usually large on both sources. Therefore it gives an attractive feature of using dominant distribution because it can maintain its distribution shape in the presence of corrupting noise.

In fig. 2, the plots of fig. 1 are depicted but with the sources being contaminated by Gaussian white noise of SNR=5dB. Notice that the dominant distributions keep their shapes much better than other ones. Also with noise effect, the optima drift away from their positions in noise-free case, which makes the adaptive search less useful.

![Fig 2: LOD in noisy environment](image)

4 Criterion for Deciding the Dominant LOD

Based on the analysis in section 3, the definition of the dominant LOD is given as follows: If the local minima with low $D_{KL}$ appear approximately along one special value of $p$ (or $q$) and hence makes two adjacent quadrants have a similar lower independence average, we call it $p$-dominant (or $q$-dominant). The global optimum is often far away from the desired one. Since one desired de-mixing matrix parameter is given directly from such LOD, the sequential algorithm is preferred [8].

Criterion for deciding the dominant LOD is based on its definition: for deciding $q$-dominant, first of all, several $p$ values are fixed such that the region $(-1,0)$ and $(0,1)$ are divided evenly in $N$ parts:

For region $(-1,0)$: $P_{-1} = \{(-1+sp),(-1+2*sp),\ldots,(-sp)\}$
For region \((0,1)\) : \(P_{0\rightarrow 1} = \{sp, 2*sp, \ldots, (1-sp)\}\)

where \(sp\) is the step-size of varying \(p\) values (obviously \(sp = 1/N\)). Therefore we have totally \(2(N-1)\) different \(p\) values. For \(p_i \in P_{-1\rightarrow 0}\) or \(p_i \in P_{0\rightarrow 1}\), calculate the K-L measurement with various \(q_j \in (-1,0)\cup(0,1)\), which is called \(KL(p_i, q_j)\).

Set \(q_{k(i)}\) as the \(q\) value at which:

\[
KL(p_i, q_{k(i)}) = \min_{q_j \in (-1,0)\cup(0,1)} KL(p_i, q_j)
\]

If

\[
\text{Max}_{i,j=1,2(N-1)} (\text{abs}(q_{k(i)} - q_{k(j)})) \leq \varepsilon
\]

where \(\varepsilon\) is a small value, then we decide that a \(q\)-dominant distribution exists. For \(p\)-dominant, the criterion is similar with several \(q\) values being fixed first.

Our proposed algorithm is based on these criteria. The mixed noisy speech signals are now first divided into a number of short-time signal-frames. Decision criteria work on every frame and only those with dominant LOD are selected for further estimation. In this way, the distributions that are heavily affected by noise will be removed. The complete algorithm is summarized in the following flow chart with details explained thereafter:

![Flow Chart](image)

**Fig 3: The Proposed Algorithm**

5 Experiments on Mandarin Chinese Noisy Speeches

Experiments have been carried out on the Mandarin Audio signal from the TDT 2 database (LDC2001s93). White noise from NOISE-ROM-0 is added to both speech sources at SNR=5dB. As we mentioned in section 3, greater difference of sources’ energy and kurtosis helps to establish the dominant distributions. Since kurtosis is
sensitive to outliers, therefore we focused on the energy difference occurrence. For Mandarin speech, there are lower-energy segments as well as higher-energy ones. When they are in the same signal-frame, it is of higher probability that dominant distribution appears because of high energy difference.

For Mandarin Chinese, the lower-energy cases include:
1) Consonants often have energy much smaller than vowels.
2) Pauses with silence for separating phases (comma and full stop are used for such pauses, and also pauses occurring when there is stress, etc)
3) Neutral tone that tend to be uttered much shorter and lighter (e.g. modal words like “[ba], [a]”; auxiliary words like “[de], [le], [zhe]”, etc)

On the other hand, the higher-energy segments might be due to:
1) Vowels comparing with consonants
2) The prominent syllables in word and the prominent words in a sentence with acute and intense utterance, for example the stress [9]

Also Mandarin is a tone language. Different tones give different energy change by short-time signal frames even in one word and may give more energy difference.

By applying the proposed algorithm, we examine the noisy Mandarin audio signal separation with one male speech and one female speech. The sampling frequency is 16000Hz. We set \( N = 3 \) for fast calculation and \( \epsilon = 0.1 \) such that there are more signal frames with dominant LOD to be selected. We found that the estimated values are sometimes affected by noise and the absolute value may be changed to larger ones. Therefore, the criterion for getting the \( p_{des} \) can be described as:

\[
p_{des} = \text{mean}(STP(1::J)), \quad q_{des} = \text{mean}(STQ(1::J))
\]  

where \( STP = \text{sort}(p_{set}) \) which means it is the vector after sorting the \( p_{set} \) from minimal absolute value to the maximum one, and so is \( STQ \) (\( p_{set} \) is originally empty vector and save all the \( p \) values into it from signal-frames with dominant distributions and so is \( q_{set} \)). Therefore the estimated de-mixing matrix is \( W_{des} = [1 \quad p_{des}; \quad q_{des} \quad 1] \). We set \( I = J = 4 \) depending on the experiments. The estimated results de-mixing matrix parameters \( p \) and \( q \) are in fig.4 for the desired de-mixing matrix=[1 \ -0.8; \ -0.8 \ 1]. The \( En(s_1) \) is calculated by \( \sum_{l=1}^{\text{Len}} (s_1(l))^2 \) and \( En(s_2) \) in the same way, with \( \text{Len} \) is the samples we used and here \( \text{Len} = 11200 \). Square window with length of 320 samples is used here for enframing to get signal frames with frame shift of 160.

In the plot of fig. 4(a)-(d), we can see the effect of sources (fig. 4(a)), noise (fig. 4(b)) and also mixing matrix (fig. 4(c)). Noise sometimes creates a larger number of frames with higher energy difference in the noisy signals but it also makes the dominant values drift a bit. In the experiment, we notice that higher energy difference occurs for sources that are voiced and unvoiced /pause correspondingly. Words with neutral tone like “[de], [le]” etc occur quite often in Mandarin even for news report, and so are the words with stress. They offer exhibitive effect to build up the dominant distributions. In fig. 4(e)-(f), we show the performance of separation by the proposed algorithm. From top to below, there is the original speech source 1 or 2, and then the
corresponding noisy signal and the lowest is the separated noisy speech signal 1 or 2 we achieved. Fig. 4(g) gives the mixtures for observation, and fig. 4(h) is the estimated $p$ and $q$ values for selected frames with dominant distribution, while for other frames, we just set the $p$ and $q$ values to be zero. We can see that the estimated $p$ and $q$ values are close to the desired de-mixing matrix parameters, and therefore the separated outputs are quite similar to the noisy sources by applying equation (8), which demonstrates that the performance of the de-mixing matrix estimation by the proposed algorithm is good.

6 Conclusion

In this paper, we propose a new algorithm for noisy speech signal separation. We show that the so-called “Dominant LOD” is more robust to noise effect and is capable of giving more accurate de-mixing matrix estimation. No prior knowledge of the noise covariance is needed and the whole PDF can be used for estimation, rather than using only higher-order cumulants, which are sensitive to outliers. Good experimental results are given to demonstrate the efficacy of the algorithm.

Furthermore, since the algorithm is based only on the “Dominant LOD”, so theoretically it may still be useful even when the noise signal is not Gaussian distributed. As shown in fig. 5, experimental results are presented for pink noise, babble noise and music instead of white noise used above, and the estimated $p$ and $q$ values, the original, noisy and separated speech signals are contrasted.

It is demonstrated that the energy difference of sources can help build up the “Dominant LOD”. Speech signals have many special characteristics of energy distribution for male or female, voiced or unvoiced, stress, tones, etc. Several examples are given to illustrate this phenomenon in time domain. Further study should be focused on detailed discussion in the time-frequency domain for more frames with “Dominant LOD”. Specific energy distributions of human speeches will be applied to obtain better separation performance.

References

Fig. 4 Experiment Results (white noise case)

(a) $	ilde{s}_1$, $s_1 + 	ilde{n}_1$, $y_1$
(b) $s_1$, $s_1 + 	ilde{n}_1$, $y_1$
(c) $s_1$, $s_1 + 	ilde{n}_1$, $y_1$

 Estimated $p$ and $q$

(a) Pink noise case
(b) Babble noise case
(c) Music noise case

Fig. 5 Experiment Results with other noises