INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Unmixing Convolutive Mixtures by Exploiting Amplitude Co-modulation: Methods and Evaluation on Mandarin Speech Recordings

Bo-Rui Chen, Huang-Yi Lee, and Yi-Wen Liu

Dept. Electrical Engineering, National Tsing Hua University, Taiwan
pbga16@hotmail.com, ian6901@yahoo.com.tw, ywliu@ee.nthu.edu.tw

Abstract

This paper presents and evaluates two frequency-domain methods for multi-channel sound source separation. The sources are assumed to couple to the microphones with unknown room responses. Independent component analysis (ICA) is applied in the frequency domain to obtain maximally independent amplitude envelopes (AEs) at every frequency. Due to the nature of ICA, the AEs across frequencies need to be *de-permuted*. To this end, we seek to assign AEs to the same source solely based on the correlation in their magnitude variation against time. The resulted time-varying spectra are inverse Fourier transformed to synthesize separated signals. Objective evaluation showed that both methods achieve a signal-to-interference ratio (SIR) that is comparable to Mazur et al (2013). In addition, we created spoken Mandarin materials and recruited age-matched subjects to perform word-by-word transcription. Results showed that, first, speech intelligibility significantly improved after unmixing. Secondly, while both methods achieved similar SIR, the subjects preferred to listen to the results that were post-processed to ensure a speech-like spectral shape; the mean opinion scores were 2.9 vs. 4.3 (out of 5) between the two methods. The present results may provide suggestions regarding deployment of the correlation-based source separation algorithms into devices with limited computational resources.

Index Terms: blind source separation (BSS), permutation problem, convolutive mixture, signal envelope, independent component analysis (ICA)

1. Introduction

Independent component analysis (ICA) has been successfully applied for source separation in various domains. For sound source separation, however, direct application of ICA in the time domain (TD) might be ineffective due to convolution of the sources with room responses. It is possible to un-mix convolutive mixtures by modeling the room response with a finite impulse response (FIR) filter, but the computation load is heavy; typically, thousands of FIR coefficients need to be estimated.

Alternatively, one may attempt to perform the separation by applying ICA in the frequency domain (FD) via the short-time Fourier transform (STFT); its plausibility is due to the fact that linear time-invariant filtering is multiplicative in the FD. However, two challenges remain to be conquered; first, the scaling factors at every frequency need to be determined, and secondly, the sources across frequencies need to be grouped because the outputs of ICA can be arbitrarily permuted. The first challenge, known as the *scaling problem*, can be solved via a minimal distortion principle [1]. The second challenge, known as the *permutation problem*, turns out to be the more difficult one and additional assumptions need to be made. In this paper, we use the term "de-permutation" to refer to the task of resolving the permutation ambiguity across frequency bins.

Existing methods for de-permutation can be categorized into two kinds [2]. Methods of the first kind utilize information derived from the unmixing matrices, such as the direction of arrivals (DOA) [3], the directivity patterns [4, 5], and time difference of arrival (TDOA) [6, 7]. These methods could be prone to errors at high frequencies due to spatial aliasing. Methods of the second kinds, instead, rely on the assumption that the correlation of the amplitude envelopes (AEs) from the same source should be high across frequencies. Clustering algorithms have thus been established for de-permutation [2, 8-11]. The two kinds of depermutation methods could also be hybridized [2] for performance optimization. In this work, we attempted to (i) investigate whether it is still possible to depermute and achieve a similar signal-to-interference ratio (SIR) as reported in [2] purely by examining the correlation between AEs (i.e., without hybridization), and (ii) to investigate whether the minimum-distortion principle provides the best solution to the scaling problem as far as user experiences are concerned. The rest of this paper is organized as follows: Sec. 2 briefly reviews the blind source separation problem. Sec. 3 describes two depermutation algorithms that we developed for this research. Sec. 4 describes the speech materials used for testing the methods, and evaluates the performance of the methods both objectively and subjectively. Discussion and conclusions are given in Sec. 5 and 6, respectively.

2. Blind source separation (BSS) for convolutive mixtures

Suppose that there are N sources and N sensors in a room and mixing of the sources can be modeled by FIR filtering,

$$\mathbf{x}(n) = \mathbf{H}(n) * \mathbf{s}(n) = \sum_l \mathbf{H}(l)\mathbf{s}(n-l) \quad (1)$$

where n denotes the time index, $\mathbf{x}(n)$ and $\mathbf{s}(n)$ are $N\times1$ vectors denoting the mixtures and the sources respectively, and $\mathbf{H}(n)$ is a sequence of $N \times N$ matrices that describes of the impulse response of the mixing channels. To separate the sources, one can attempt to use finite impulse-response (FIR) filters [13]:

$$\mathbf{y}(n) = \mathbf{W}(n) * \mathbf{x}(n) = \sum_{l=0}^{L-1} \mathbf{W}(l)\mathbf{x}(n-l) \quad (2)$$

where $\mathbf{W}(n)$ is a sequence of unmixing matrices to be estimated, and L is the length of the filters. Identification $\mathbf{W}(n)$ involves blind de-convolution and is a difficult task. In this research, we choose to transform the problem into the frequency domain. Equation (1) can be written as follows,

$$\mathbf{X}(k,q) = \mathbf{H}_{\mathrm{f}}(k)\mathbf{S}(k,q), \quad (3)$$

where k is the frequency index, $\mathbf{H}_{\mathrm{f}}(k)$ denotes the discrete

http://dx.doi.org/10.21437/Interspeech.2017-1227

Fourier transform of $\mathbf{H}(n)$, and q denotes the frame number. At every frequency k, the complex-valued ICA [12] can be applied over a sequence of frames $q = 1, \ldots, Q$ to identify an unmixing matrix $\mathbf{W}_{\mathrm{f}}(k)$ of size $N \times N$ that best separates the mixed signals; we have

$$\mathbf{Y}(k,q) = \mathbf{W}_{\mathrm{f}}(k)\mathbf{X}(k,q), \quad (4)$$

where the resulting $\mathbf{Y}(k,q)$ can be regarded as if it contains N independent outputs of length Q at every frequency k. However, ICA algorithms typically involve random initialization, so the resulting matrix $\mathbf{W}_{\mathrm{f}}(k)$ is ambiguous against scaling and permutation of the rows. Extra criteria need to be adopted to resolve the ambiguities.

3. The proposed algorithms

In this section, we describe two de-permutation methods that have been developed for this research. For discussion purposes, an *envelope correlation coefficient* (ECC) between frequencies *k* and *l* is defined as follows [2],

$$\rho_{ij}(k,l) = \frac{\sum_{q=1}^{Q} V_i(k,q)V_j(l,q)}{\sqrt{\sum_{q=1}^{Q} V_i^2(k,q)}\cdot\sqrt{\sum_{q=1}^{Q} V_j^2(l,q)}} \quad (5)$$

where $V_i(k,:) = |Y_i(k,:)|$ is the AE for the *i*th separated source produced by complex-valued ICA at frequency *k* (the ":" symbol of Matlab is adopted here to denote all the elements in a row). In the rest of the paper we focus on separation of two sources recorded by two channels. To determine the correct permutation between bin *k* and *l*, two possible alignment options exist,

(i) Align $V_1(k,:)$ with $V_1(l,:)$, and $V_2(k,:)$ with $V_2(l,:)$.

(ii) Align $V_1(k,:)$ with $V_2(l,:)$, and $V_2(k,:)$ with $V_1(l,:)$.

Ideally, the alignment can be determined based on the following ratio [2]:

$$\gamma(k,l) = \frac{\rho_{11}(k,l)+\rho_{22}(k,l)}{\rho_{12}(k,l)+\rho_{21}(k,l)}, \quad (6)$$

and if $\gamma(k,l) > 1$, option **(i)** is preferred and vice versa. However, in practice it often happens that $\gamma(k,l) \cong 1$ at adjacent frequencies and how best to align the AEs becomes uncertain.

3.1. Method A: "clustering first"

In this method, the entire frequency range is clustered into segments first so that within each segment we could perform permutation to ensure that the AEs of adjacent frequency bins all satisfy the following criterion,

$$\min(\rho_{11}(k,k+1),\ \rho_{22}(k,k+1)) > \max\big(\rho_{12}(k,k+1), \rho_{21}(k,k+1)\big) + d, \quad (7)$$

where we empiricially chose d = 0.2. Note that (7) is a stricter criterion than requiring $\gamma > 1$.

Next, we identify all the segments {seg1, …, segK} that are wider than 150 Hz. For each segment narrower than 150 Hz, we align it with the nearest segment among {seg1, …, segK} at a lower frequency based on the following rule,

$$\sum_{p=0}^{P-1}\sum_{r=0}^{R-1}\gamma(k_{\max}-p, l_{\min}+r) \gtrless 1? \quad (8)$$

where $k_{\max}$ is the highest frequency bin of the wide segment, $l_{\min}$ is the lowest frequency bin of the narrow segment, and *P* and *R* denotes the number of frequencies to be considered in this calculation. Empirically, we set both *P* and *R* to be equivalent to the bandwidth of 100 Hz. Finally, (8) is used repeatedly to align segments that now become all wider than 150 Hz to complete de-permutation. Note that Eq. (8) differs from Eq. (7) in the sense that it does not contain a margin d. In other words, the de-permutation decision is final.

3.2. Method B: zone expansion from an anchor chunk

In this method, we first exhaustively considered all chunks of 5 consecutive frequency bins and calculate depermutation confidence scores for each chunk. For the convenience of discussion, let "1" denote the situation when switching is necessary, and "0" denote otherwise. Let $b_1b_2b_3b_4 \in \mathcal{B} = \{0,1\}^4$ denote a binary sequence of length 4 corresponding to a possible choice of de-permutation; for instance, the choice of permutation depicted by Fig. 1 is denoted by {0011}. Then, a score function corresponding to this choice is calculated as follows,

$$S_{\mathrm{Ch1}}(\{0011\}) = \rho_{11}(f_1,f_2)\rho_{11}(f_2,f_3)\rho_{12}(f_3,f_4)\rho_{21}(f_4,f_5),$$

where $f_1, \ldots, f_K$ are K consecutive frequencies. The score for any other sequence in $\mathcal{B} = \{0,1\}^4$ is defined similarly, and the same procedure is applied to the other channel to obtain scores $S_{\mathrm{Ch2}}(b)$ for all sequences $b \in \mathcal{B}$. The final confidence score S_{tot} for any $(b_1, b_2) \in \mathcal{B}\times\mathcal{B}$ is defined as follows,

$$S_{tot}(b_1,b_2) = S_{\mathrm{Ch1}}(b_1) + S_{\mathrm{Ch2}}(b_2). \quad (9)$$

The chunk that has the largest difference between the highest and the lowest scores among all 16x16 choices in $\mathcal{B}\times\mathcal{B}$ is regarded as an *anchor*, because we have a rather high certainty about the correct permutation for that chunk being the choice that gives the highest score. The anchor chunk is hereafter considered de-permuted. Subsequently, we expand the de-permuted zone to its left and to its right by looking at one more frequency each time. Let us take f_6 as an example. We determine that the correct permutation between f_5 and f_6 is "0" if

$$\prod_{m=1}^{5}\rho_{11}(f_6, f_{6-m}) + \prod_{m=1}^{5}\rho_{22}(f_6, f_{6-m}) > \prod_{m=1}^{5}\rho_{12}(f_6, f_{6-m}) + \prod_{m=1}^{5}\rho_{21}(f_6, f_{6-m}).$$

The same procedure repeats until permutation is determined for all frequency bins.

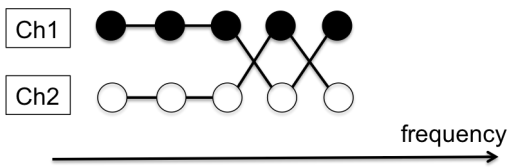

Fig. 1: *a possible choice of permutation for 5 consecutive frequencies, denoted as* {0011}.

3.3. Post processing and inverse Fourier transform

The scaling problem has been regarded as a less challenging problem than de-permutation. Nevertheless, we empirically found that the method proposed in [1] tends to result in a rather flat spectrum, which contains excessive high intensity at high frequencies and reduces the comfort of listening. Therefore, instead of applying the minimal distortion principle [1], we devised a heuristic post-processing method to deal with the scaling problem. First, a Gaussian mixture model is compared against that of the mixture signals, and the difference in the mean is compensated so that the resulting

spectral envelope becomes more speech-like. This is equivalent to multiplying a different scaling factor across all frequencies and for each channel,

$$\mathbf{Y}'(k,:) = \mathrm{diag}(\lambda_1, \lambda_2)\mathbf{Y}(k,:),$$

where $\mathrm{diag}(\lambda_1, \lambda_2)$ denotes a diagonal matrix of size 2×2, and scaling factors λ_1 and λ_2 both vary against frequency k. This technique is hereafter referred to as *speech-shaped filtering*.

Afterwards, ${Y_1}'(k,q)$ and ${Y_2}'(k,q)$ are subject to inverse Fourier transform respectively,

$$y_i^{(q)}[n] = \frac{1}{M}\sum_k {Y_i}'(k,q)\exp\left(jk\frac{2\pi}{M}n\right), i = 1,2, \quad (10)$$

where M denotes the number of samples in a frame. Finally, time-domain signals are synthesized by summing up $y_i^{(q)}[n]$ in an overlap-add manner as the frame number q proceeds. In this research, we set $M = 2048$ samples and the sampling rate is 16 kHz. The original signals are all 5 second long so the number of frames Q is 39.

4. Materials and results

For objective evaluation, we used the mixed speech signals materials available at [13] as the test materials. The output of de-permutation method A was post-processed by a minimum-distortion principle [1], while the output of method B was post-processed by speech-shaped filtering as described in Sec. 3.3. The average signal-to-interference ratio (SIR) for both methods and the computation time are listed in Table 1 – the execution time was measured by running Matlab implementation on a notebook with Intel Core i5 CPU at 2.8 GHz clock rate.

Table 1: *Objective comparison of separation performance. SIR = Signal to interference ratio.*

De-permutation Method	SIR (dB)	Computation time
Method A	20.4	9.8 sec
Method B	20.1	7.0 sec
Mazur et al. [2]	17.3	14.5 sec

While the results in Table 1 might seem to suggest that both methods achieve a better performance than reported in [2], the difference is probably not significant because we found that the SIR strongly depends on the materials. This will be further discussed below when reporting the results of unmixing Mandarin speech.

The spoken Mandarin materials consist of three mixtures. One is a mixture of two female voices, another one is a mixture of a female voice and a male voice, and the third one is a mixture of two male voices. These mixtures were obtained by simultaneously playing one clean signal from each of two loudspeakers (to emulate two talking persons) and recording the signals back by a pair of microphones in an office.

The microphones were placed 15 cm apart from each other, the loudspeakers were placed at 40 cm from the midpoint between the microphones with an azimuth angle of ±60 degrees, respectively. The microphones and the loudspeakers were all placed on a desk that was enclosed by cubicle boards on three sides. We expected the recording to be subject to reverberation to a certain extent. The recording was conducted while keeping the background as quiet as possible.

Table 2: *Objective performance of the proposed methods using Mandarin speech mixtures as the test materials. F = Female, and M = Male.*

Materials	SIR (dB) Method A	SIR (dB) Method B
F+F	6.4	7.1
F+M	15.6	19.6
M+M	13.2	13.9

Table 2 shows the SIR achieved by both of the proposed methods. It appears that the results vary significantly from one mixture to another. The mixtures and the unmixing results produced by Method B in .wav format can be found in supplementary materials, under the file names `{FF,FM,MM}-{Mix,ICA}-{1,2}.wav`. (We found a minor programming error in calculating the location of the anchor chunk in Method B. Since then, the programming error has been fixed, and the supplementary materials have been updated accordingly. However, the subjective listening test was completed prior to our discovery of the bug. After comparing to the unmixing results obtained before and after fixing the error, we decided that the sound quality does not change significantly as much as speech intelligibility is concerned. Therefore, we choose to report the subjective evaluation results that were obtained using the old listening materials, which can also be found in supplementary materials under the file names `{FF,FM,MM}-Old-{1,2}.wav`).

To investigate if the proposed methods enhance the speech in terms of intelligibility, we recruited a total of 66 subjects, all residents of Taiwan, to perform word-by-word transcription. Because that SIR for Method B seemed to be higher than that of Method A (Table 2), we decided to focus on Method B in this part of the study. The subjects were separated into two groups that are matched in both their age and their educational background (Table 3). The subjects in the first group were asked to transcribe by listening only to the speech mixtures, and the subjects in the second group listened to the unmixing results produced by Method B. The subjects all listened to the test materials via the same headphone connecting to the same computer, and they were allowed to adjust the volume and listen to the samples as many times as they would. The word transcription accuracy is shown in Table 4; Group 2 consistently performed better than group 1, which validated that the method worked well for unmixing these particular test materials.

Table 3: *Constituency of subjects in both groups*

Group 1		
Education	**N**	**Age**
College and above	27	25.2 ± 3.7
High school	5	51.8 ± 6.5
Elementary school	1	12
Group 2		
College and above	27	25.0 ± 2.9
High school	5	54.4 ± 6.0
Elementary school	1	11

Table 4: *Comparison of word transcription accuracy*

Materials	Group 1	Group 2
F+F	45.3%	86.1%
F+M	71.2%	97.8%
M+M	52.9%	94.8%

5. Discussion

The two proposed methods differ mainly in the strategy of de-permutation. Method A is more conservative in the sense that adjacent bins within a "cluster" must be aligned with high certainty; this is ensured by setting d at a rather high margin of 0.2 in Eq. (7). The "clustering first" strategy described in Method A is inspired by [2], but they differ in several ways; first, in [2], all bins within a cluster were ensured to be *pairwise* correctly aligned. However, in [2] the margin d was zero, which is not as strict as in Method A. Finally, in [2] the clustering step was followed by a merging strategy that took the average time-difference of arrival (TDOA) at every cluster into consideration, while in Method A we seek to de-permute solely based on the ECC across frequencies (Eq. 5).

Method B does not require clustering, and hence the computation cost is lower. Nevertheless, it achieves comparable performance in terms of the SIR when compared against Method A. It is worth reporting that we also asked the subjects to judge effectiveness of unmixing by giving a clarity score on a scale of 1 to 5 (1.0 = "cannot recognize the words at all", 2.0 = "tolerable", 3.0 = "fair", 4.0 = "good", 5.0 = "very good"). Somewhat remarkably, the mean score was 4.3 ± 0.56 for Method B (mean $\pm$ 1 st. dev.) and 2.9 ± 0.92 for Method A. So Method B produced results that sounded superior to those of Method A in this regard even though Method A follows a "minimum distortion principle" when dealing with the scaling problem (and though we later found a programming bug in Method B). It remains uncertain whether the subjects' overall preference to Method B suggests that its de-permutation is superior to Method A, or it is the speech-shaping filter that caused the participating subjects to prefer to it. Follow-up studies are necessary to tease apart which of the afore-mentioned factors are most crucial. Finally but not the least important, the simplicity and rigor of the algorithms can also be improved, because parameters in the current methods were undeniably tuned manually by trial and error before we recruited subjects to conduct listening tests; it may also be worthy to consider and compare against a vastly different approach that circumvents the permutation problem by constrained ICA [14].

6. Conclusions

In this study, we developed two de-permutation strategies to perform BSS in the frequency domain via ICA. Both of them achieved similar levels of SIR when compared against a state of the art [2]. When the two methods are compared against each other, we found that de-permutation might not require clustering the frequency bins first; a sequential alignment might just do the job equally well as long as an "anchor" chunk can be identified at the beginning. User feedback partially supported that the sequential alignment produced no inferior quality than using the more conservative "clustering first" strategy. Further studies are warranted to pinpoint which of the factors being adjusted in this research are crucial for BSS. Nevertheless, the sequential alignment (Method B) should be preferable if computational cost is a main concern. This may provide some guidelines if one ever needs to deploy similar BSS algorithms to devices with limited computational power, such as any wearable assistive hearing apparatus.

7. Acknowledgements

The authors would like to thank the Ministry of Science and Technology of Taiwan for supporting this research under grant No. 103-2221-E-007-085-MY2. The authors thank four anonymous reviewers for providing valuable critiques.

8. References

[1] K. Matsuoka, "Minimal distortion principle for blind source separation," in *Proc. 41st Soc. Instruments and Control Engineers Annual Conference*, 2002, vol. 4, pp. 2138–2143.

[2] R. Mazur, J. O. Jungmann, and A. Mertins, "A new clustering approach for solving the permutation problem in convolutive blind source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[3] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, Sep. 2004.

[4] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP),* 2000, vol. 5, pp. 3140–3143.

[5] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proc. IEEE ICASSP*, 2002, vol. 1, pp. I–881-884.

[6] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components with estimating propagation model parameters in frequency-domain blind source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, July 2007.

[7] F. Nesta and M. Omologo, "Approximated kernel density estimation for multiple TDOA detection," in *Proc. IEEE ICASSP*, May 2011, pp. 149 –152.

[8] S. Ikeda and N. Murata, "A method of blind separation based on temporal structure of signals." in *Proc. Int. Conf. Neural Information Processing*, Oct. 1998, pp. 737–742.

[9] V. G. Reju, S. N. Koh and I. Y. Soon "A robust correlation method for solving permutation problem in frequency domain blind source separation of speech signals," *Proc. IEEE Asian Pacific Conf. Circuits and Systems*, Dec. 2006, pp.1893 -1896.

[10] K. Rahbar and J. P. Reilly, "A frequency domain method for blind source separation of convolutive audio mixtures," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 832–844, Sept. 2005.

[11] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *Proc. IEEE Int. Symp. Circuits and Systems*, 2007, pp. 3247–3250.

[12] E. Bingham and A. Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," *Int. J. Neural Systems*, vol. 10, no. 1, pp. 1-8, 2000.

[13] http://www.kecl.ntt.co.jp/icl/signal/sawada/demo/bss2to4/index.html

[14] F. Nesta and M. Matassoni, "Robust automatic speech recognition through on-line semi blind source extraction," in *Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, Sep. 2011, pp. 18-23.