STEREO SOURCE SEPARATION IN THE FREQUENCY DOMAIN: SOLVING THE PERMUTATION PROBLEM BY A SLIDING K-MEANS METHOD

Bang-Yin Chen, Tzu-Chi Liu, Yi-Wen Liu

Dept. Electrical Engineering, National Tsing Hua University, Taiwan

ABSTRACT

Blind source separation (BSS) has been widely utilized for recovering a set of source signals from their mixtures. When the mixture is convolutive, source separation can be solved in the frequency domain but involves several challenges including the scaling uncertainty and the permutation indeterminacy. This paper presents a sliding k-means algorithm to handle the permutation problem. Experiments were conducted by playing the source files to a pair of loudspeakers and obtaining the mixture by microphones. Objective indices are then defined to evaluate the separation performance based on the actual frequency responses. Results have shown that the standard k-means method alone can consistently achieve $> 90.5\%$ permutation accuracy in different parameter settings. After introducing the proposed sliding process, the permutation accuracy further rises. Compared to a previous de-permutation method [1], the present method has a more stable performance against parameter variations in terms of its permutation accuracy and signal-to-interference ratio (SIR).

Index Terms— blind source separation (BSS), permutation problem, independent component analysis (ICA)

1. INTRODUCTION

Blind source separation (BSS) is a task to reconstruct individual sound sources from their mixtures which are synchronously received by a set of microphones [2]. Independent component analysis (ICA) [3] can serve as a statistical tool for solving this problem [4, 5]. In an actual room, signals are mixed convolutively with reverberations, demanding a matrix of FIR filters to be estimated [6]. For simplification, the signals can be short-time Fourier transformed and processed in the time-frequency domain [7, 8]. Thus, complexed-valued ICA for the instantaneous mixtures can be applied at each frequency bin [9]. However, the permutation indeterminacy emerges after ICA and the components from the same source signal need to be somehow aligned across different frequencies.

The alignment problem has been addressed previously based on an assumption that the correlation between the amplitude envelopes of adjacent bins should be higher from the same source than from different sources [10, 1]. However, the rules for de-permutation in previous methods are typically designed heuristically [1], could become complicated and hard to manage, and are perhaps prone to single-frequency error. Therefore, we aimed to mend this problem by partitioning frequency bins into several batches and then perform de-permutation based on clustering by the k-means algorithm.

To evaluate and compare the performance of different methods, we objectively measured the frequency response of mixing and defined an index for evaluating the performance of de-permutation. Ideally, the mixing matrix and the un-mixing matrix should be inverse to each other. One could measure the room responses as the ground truth and calculate whether the product of the mixing matrix and the un-mixing matrix concentrates on the diagonal. To quantify the degree of diagonal concentration, the concept of Pearson correlation coefficient was adopted [11]. The organization of the rest of this paper is as follows: Sec. 2 gives a brief review of the blind source separation. Sec. 3 introduces the proposed de-permutation algorithm. Sec. 4 describes the testing materials and the objective index to help evaluate the results. In Sec. 5 and 6, discussions and conclusions are made, respectively.

2. BLIND SOURCE SEPARATION (BSS) FOR CONVOLUTIVE MIXTURES

Given a pair of sources $\boldsymbol{s} = (s_1, s_2)^T$ and a pair of microphones in a room, the convolutive mixtures $\boldsymbol{x} = (x_1, x_2)^T$ can be modeled as

$$\boldsymbol{x}(n) = \boldsymbol{h}(n) * \boldsymbol{s}(n) = \sum_{l=0}^{L-1} \boldsymbol{h}(l)\boldsymbol{s}(n-l), \qquad (1)$$

where n denotes the time index, L is the presumed FIR filter length, and the sources $\boldsymbol{s}(n)$ are convolved with the impulse response of the room $\boldsymbol{h}(n) \in \mathbb{R}^{2\times 2}$. The goal is to find the un-mixing matrices $\boldsymbol{w}(n)$ so that the output

$$\boldsymbol{y}(n) = \boldsymbol{w}(n) * \boldsymbol{x}(n) = \sum_{l=0}^{L-1} \boldsymbol{w}(l)\boldsymbol{x}(n-l) \qquad (2)$$

approximates $\boldsymbol{s}(n)$ as much as possible. Since blind deconvolution of $\boldsymbol{w}(n)$ is a tough task, the problem is Fourier

transformed into the frequency domain in this research. Equation (1) can be written as

$$\boldsymbol{X}(k,q) = \boldsymbol{H}(k)\boldsymbol{S}(k,q), \tag{3}$$

where k denotes the bin number, q denotes the frame number, and $\boldsymbol{H}(k)$ is the discrete Fourier transform of $\boldsymbol{h}(n)$. Then, the complex-valued ICA [9] is applied over a sequence of frames at each frequency bin k. Thus, we obtain

$$\boldsymbol{Y}(k,q) = \boldsymbol{W}(k)\boldsymbol{X}(k,q) = \begin{bmatrix} Y_1(k,q) \\ Y_2(k,q) \end{bmatrix}, \tag{4}$$

where $Y_1(k,:)$ and $Y_2(k,:)$ are independent components (here, the symbol ':' denotes all elements in a row). However, the permutation indeterminacy comes after ICA and thus it requires extra efforts to resolve.

3. THE PROPOSED ALGORITHMS

In this section, the proposed sliding k-means method for handling the permutation problem is introduced. Instead of calculating the correlation among frequency bins and linking the frequency bins successively as suggested in [1], frequency bins are first partitioned into overlapping batches and bins in every batch is clustered by the k-means method.

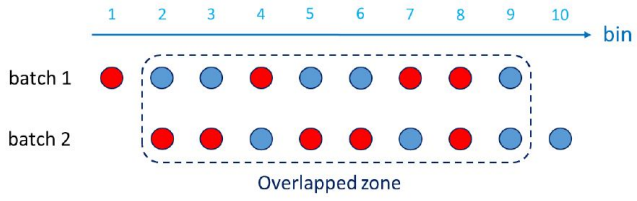
Assume that L frequency bins are divided into N_{bat} batches for both channels, where each batch contains B neighboring frequency bins and adjacent batches are overlapped with B_{ove} frequency bins. Thus, we can have a hop size of $B_{\text{hop}} = B - B_{\text{ove}}$ and the number of batches can be calculated as $N_{\text{bat}} = \lceil \frac{L - B_{\text{ove}}}{B_{\text{hop}}} \rceil$. Then the k-means method is applied to all the batches one by one, so the proposed method is referred to as *sliding k-means*.

For the de-permutation within a batch, the amplitude envelope $S_j(k) = |Y_j(k,:)|$ is treated as a vector for the j-th source at the k-th frequency bin in the batch, its length being the number of frames. At the beginning, a pair of random frequency bins are initialized as centroids while the rest of the frequency bins are assigned to one of the clusters based on their 1-norm distance to the centroids. The objective function is defined as

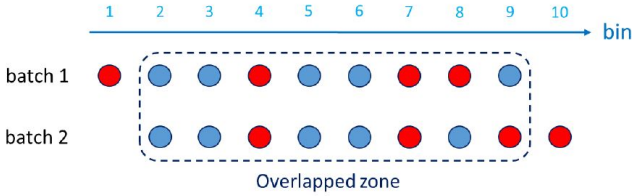
$$J = \underset{c_j, j=1,2}{\arg\min} \sum_{j=1}^{2} \sum_{k=1}^{B} ||S_j(k) - c_j||, \tag{5}$$

where c_j is the centroid for j-th source. Once all the members in each cluster are determined, the locations of the centroids can be updated. Subsequently, the 1-norm distances between centroids and the members in each cluster are updated again. The iteration process is terminated until the sum of square of the distances converges or the iteration number reaches 30.

After applying the k-means algorithm in every batch, it is possible that results at the same frequency bin may be assigned to different clusters in different batches. Thus, frequency bins in the overlapped zone are checked first to see whether the classification results are highly similar but oppositely labeled; Fig. 1(a) illustrates an example, where the two channels are colored blue and red, respectively. If so, the labels need to be flipped so that the present batch and the previous batch reach unanimous decision on more than half of the frequency bins; by doing so, the frequency bins become consistently labeled as shown in Fig. 1(b).



(a) An oppositely labeled batch.



(b) A consistently labeled batch.

Fig. 1. Illustration of the batch labeling.

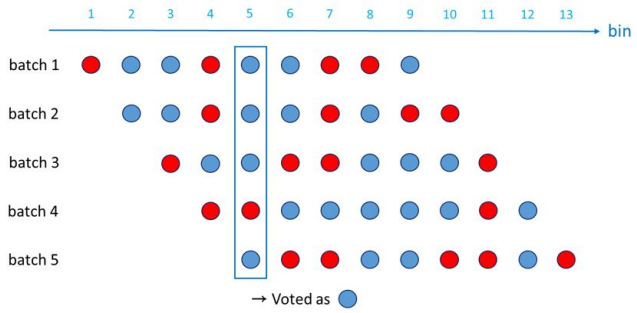


Fig. 2. Illustration of the cluster voting for each frequency bin, respectively.

Afterwards, the final clustering decision for each frequency bin (the channel it belongs to) is determined by voting the dominant labels across different batches as illustrated in Fig. 2.

4. MATERIALS AND EVALUATION OF RESULTS

To test the performance of the proposed method, 4 Mandarian songs were pre-recorded by two male and two female singers, respectively. The mixtures were obtained by simultaneously playing individual clear signals by loudspeakers and recording the signals back by a pair of microphones in an actual room as illustrated in Fig. 3. We prepared three differ-

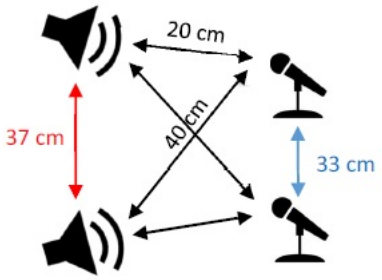


Fig. 3. Equipment setting for the experiments.

ent conditions of mixtures: male+male, female+female, and male+female.

Ideally, if the de-permutation is done correctly, the product of the mixing matrix $\boldsymbol{H}(k)$ and the un-mixing matrix $\boldsymbol{W}(k)$ should be highly concentrated on the diagonal. For evaluation purposes, the actual coupling responses was measured in the room and the mixing matrix was obtained as the ground truth. The un-mixing matrix can be calculated after going through ICA and solving the scaling problem and the permutation problem. Then, the actual product $\boldsymbol{P}(k)$ is defined as follows:

$$\boldsymbol{P}(k) = \boldsymbol{H}(k) \times \boldsymbol{W}(k). \tag{6}$$

For the convenience of calculation, the magnitude part of $\boldsymbol{P}(k)$, denoted as $p_{ij}(k)$, is taken as

$$p_{ij}(k) = \sqrt{\boldsymbol{P}_{ij}(k) \times \boldsymbol{P}^*_{ij}(k)}, \text{ where } \begin{cases} i = 1, 2. \\ j = 1, 2. \end{cases} \tag{7}$$

Then, the modified sample Pearson correlation coefficient $\gamma(k)$ is introduced to evaluate the diagonal concentration, defined as follows:

$$\gamma(k) = \frac{S_I S_{Ixy} - S_{Ix} S_{Iy}}{\sqrt{S_I S_{Ixx} - (S_{Ix})^2}\sqrt{S_I S_{Iyy} - (S_{Iy})^2}}, \tag{8}$$

where $S_I = p_{11} + p_{12} + p_{21} + p_{22}, S_{Ix} = p_{11} + p_{12} + 2p_{21} + 2p_{22}, S_{Iy} = p_{11} + 2p_{12} + p_{21} + 2p_{22}, S_{Ixx} = p_{11} + p_{12} + 4p_{21} + 4p_{22}, S_{Iyy} = p_{11} + 4p_{12} + p_{21} + 4p_{22}$, and $S_{Ixy} = p_{11} + 2p_{12} + 2p_{21} + 4p_{22}$. At any frequency f, if $\gamma(k)$ is close to 1, it indicates the successful separation and the correct de-permutation of the resulting signals at that frequency bin because $\boldsymbol{P}(k)$ is close to a diagonal matrix; if $\gamma(k)$ is close to -1, it indicates successful separation yet the wrong permutation of the signals. Finally, if $\gamma(k)$ is close to 0, it indicates that ICA fails and there has been little confidence to de-permute the signals correctly.

Fig. 4 shows two examples of the the above-defined score across different frequencies. To evaluate the overall accuracy of de-permutation, a permutation accuracy r_{acc} is defined as follow,

$$r_{acc} = \frac{\max\{N_p, N_n\}}{L}, \tag{9}$$

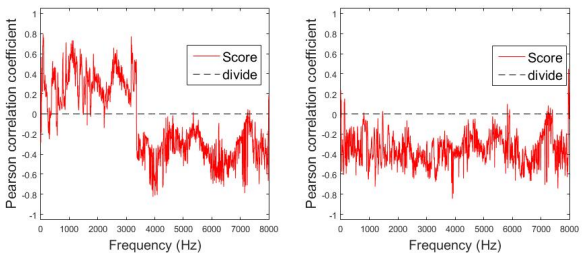


Fig. 4. Illustrations of the score $\gamma(k)$ across different frequencies. Note that, for the example shown on the left panel, a single-frequency error must have occurred between 3 to 4 kHz and it ruins the overall result of de-permutation.

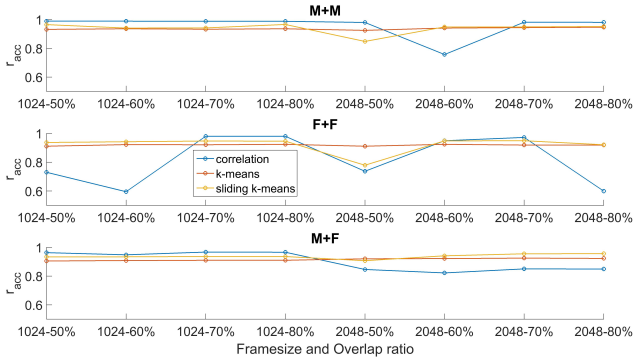


Fig. 5. Comparison of permutation accuracy of different methods on different materials ('M' for the male, 'F' for the female) under various parameter settings.

where N_p represents the number of frequency bins with positive scores (i.e., $\gamma(k) > 0$) and N_n represents the number of frequency bins with negative scores (i.e., $\gamma(k) < 0$).

With the definition of the scoring system, the sliding k-means method in this paper can be compared with the method [1] that is based on calculation of correlation between adjacent frequency bins.

As shown in Fig. 5, the permutation accuracies are presented across three methods for comparsion. To evaluate the stability of the system, we examine the permutation accuracy across different settings of system parameters, including the frame size and the overlap ratio. In Fig. 5, the correlation method has the highest permutation accuracy in 15/24 of the cases, but falls significantly to the lowest in the other 9 cases. In contrast, the performance of the k-means method is stable, though it is not the best in almost all conditions (23/24) except one point (F+F, 2048-50%) in Fig. 5. The sliding k-means method is more stable than the correlation method, and its accuracy is higher than the k-means method in 18/24 of the cases.

Besides the permutation accuracies, the average signal-to-interference ratios (SIR) [12] before and after source separa-

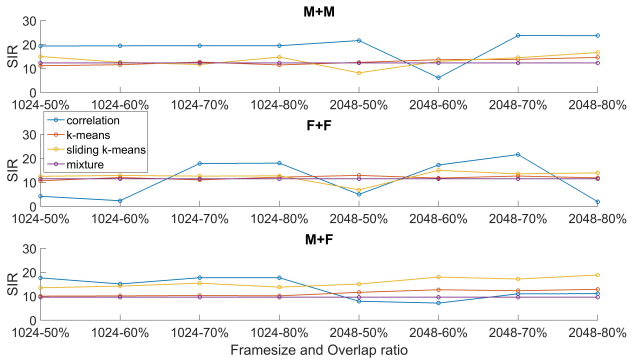


Fig. 6. Signal-to-interference ratio obtained by different methods on different materials (M+M, F+F, M+F) with different parameter settings. The SIR in the original mixtures (i.e., before source separation) is also shown for comparison.

tion by the three methods are compared in Fig. 6. The results are similar to those shown in Fig. 5. Note that, when the correlation method does not perform well, it can even reduce the SIR (in 6/24 cases) after "source separation". In contrast, the k-means method consistently improved the SIR in all 24 cases. Although the SIR for the sliding k-means method reduces the SIR in 2 cases (M+M, 2048-50% and F+F, 2048-50%), it peforms better than the plain k-means method in 21/24 cases.

5. DISCUSSIONS

The correlation method can achieve high permutation accuracy if the parameters are selected properly. However, a single-frequency error (see the left panel of Fig. 4) may ruin the de-permutation results. The results in Fig. 5 suggest that the correlation method might be unstable at some parameter settings and end up with worse permutation accuracies than those of the plain k-means algorithm, which are consistently over 90.5%. Compared to the correlation method, the average improvement in permutation accuracy is 3.2% and 4.1% for the plain k-means and the sliding k-means method, respectively.

Essentially, the reason why the k-means method is more stable than the correlation method is because that the k-means algorithm does not solely cluster two adjacent frequency bins, instead, it considers all the frequency bins together. However, if the frequency range is chosen too broadly, the pattern of the amplitude envelopes $Y_j(k,:)$ may diversify even when retrieved from the same source. Therefore, the sliding process is introduced to ensure that an appropriate frequency range is chosen.

Figure 7 illustrates the sliding k-means method after reducing the amplitude envelopes to 3 dimensions via principal component analysis. The features do spread out and cluster in

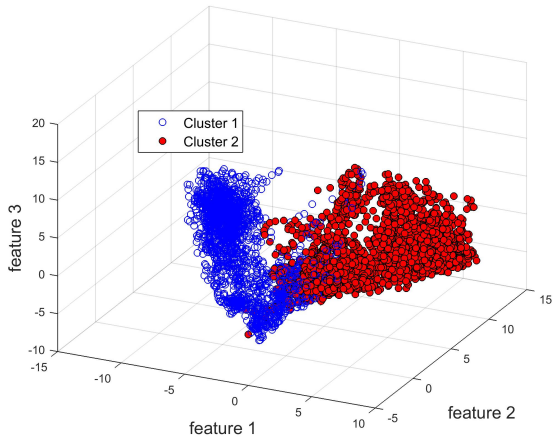


Fig. 7. A 3-dimensional visualization of results of solving the permutation problem by the sliding k-means method.

3D but there is an ambiguous zone where two clusters meet and overlap each other. Empirically, we have found that the frequency bins in the ambiguous zone have a high risk to be wrongly clustered, and this is a potential place for improvement.

6. CONCLUSIONS

We have proposed a new method for solving the permutation problem in frequency-domain BSS. Compared with one previous de-permutation method based on correlation between the amplitude envelopes of adjacent bins [1], the present methods based on the k-means algorithm are more stable when encountering different parametric settings. The above conclusion has been made based on two performance metrics: the first is a *permutation accuracy* obtained by multiplying the derived un-mixing matrix with the actual mixing matrix measured in the recording room; the second is the SIR. Results have shown that the proposed method can handle the permutation problem with higher robustness compared to the previous method [1] because it is more immune to single-frequency errors.

Currently, the algorithms have been tested on singing-voice mixtures of musically unrelated melodies. It would be of interest to test the present method on mixtures of harmoniously sung sources, such as in an *a cappella* setting. We expect this to be more challenging than unmixing the present materials and future research is warranted.

7. ACKNOWLEDGEMENTS

This research is supported by the Ministry of Science and Technology of Taiwan (MOST 108-2634-F-007-003).

8. REFERENCES

[1] B.-R. Chen, H.-Y. Lee, and Y.-W. Liu, "Unmixing convolutive mixtures by exploiting amplitude comodulation: Methods and evaluation on mandarin speech recordings," in *Proc. Interspeech 2017*, Aug. 2017, pp. 1934–1937.

[2] J. F. Cardoso, "Blind signal separation: statistical principles," in *Proceedings of the IEEE*, Oct. 1998, vol. 86(10), pp. 2009–2025.

[3] S. Choi, A. Cichocki, H.-M. Park, and S.-Y. Lee, "Blind source separation and independent component analysis: A review," *Neural Information Processing-Letters and Reviews*, vol. 6(1), pp. 1–57, Jan. 2005.

[4] J. F. Cardoso and A. Soulomiac, "Blind beamforming for nongaussian signals," *Neural Networks*, vol. 13(4-5), pp. 411–430, 2000.

[5] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," in *IEE proceedings-F*, Dec. 1993, vol. 140(6), pp. 362–370.

[6] S. Araki, S. Makino, R. Mukai, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11(2), pp. 109–116, 2003.

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

[8] 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(5), pp. 832–844, Sept. 2005.

[9] E. Bingham and A. Hyvrinen, "A fast fixed-point algorithm for independent component analysis of complex-valued signals," *Int. J. Neural Systems*, vol. 10(1), pp. 1–8, Jan. 2000.

[10] 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*, Oct. 2013, pp. 1–4.

[11] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*, pp. 1–4. Springer, Berlin, Heidelberg, 2009.

[12] E. Vincent, R. Gribonval, and C. Fvotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14(4), pp. 1462–1469, 2006.