Solving permutations in frequency-domain for blind separation of an arbitrary number of speech sources

Iván Durán-Díaz, Auxiliadora Sarmiento, Sergio Cruces, et al.

Citation: The Journal of the Acoustical Society of America **131**, EL139 (2012); doi: 10.1121/1.3678657
View online: https://doi.org/10.1121/1.3678657
View Table of Contents: https://asa.scitation.org/toc/jas/131/2
Published by the Acoustical Society of America

Solving permutations in frequency-domain for blind separation of an arbitrary number of speech sources

Iván Durán-Díaz, Auxiliadora Sarmiento, Sergio Cruces, and Pablo Aguilera
Signal Theory and Communications Department, University of Seville, Camino de los Descubrimientos S/N, 41092, Seville, Spain
duran@us.es, sarmiento@us.es, sergio@us.es, paguilera@us.es

Abstract: Blind separation of speech sources in reverberant environments is usually performed in the time-frequency domain, which gives rise to the permutation problem: the different ordering of estimated sources for different frequency components. A two-stage method to solve permutations with an arbitrary number of sources is proposed. The suggested procedure is based on the spectral consistency of the sources. At the first stage frequency bins are compared with each other, while at the second stage the neighboring frequencies are emphasized. Experiments for perfect separation situations and for live recordings show that the proposed method improves the results of existing approaches.

PACS numbers: 43.60.Pt, 43.60.Gk, 43.60.Np [CG]
Date Received: October 4, 2011 **Date Accepted:** December 16, 2011

1. Introduction

Blind separation of speech signals is a challenging problem whose solution strongly depends on the mixing conditions and on the assumed hypotheses. In reverberant environments it is usual to work in the time-frequency domain, thus decoupling the problem into a set of instantaneous problems with complex-valued sources.[1] In spite of this advantage, an obstacle arises, the permutation problem, consisting of the fact that the ordering of the estimated sources does not coincide for different frequency bins. Methods for solving the permutation problem can be grouped into two categories:[1] methods exploiting the consistency of the filters coefficients and methods exploiting the consistency of the spectrum of the sources. The second group of methods is based on the inter-frequency coherence of the spectrum of speech signals, known as the amplitude modulation correlation property.[2] In this work we introduce a new method based on this property to solve the permutation problem in a general reverberant environment with N sources and N sensors. Unlike the best of the state of the art methods,[3] which have been conceived for the 2×2 case, but not directly extended to the $N \times N$ case, in this work N can be greater than 2.

The paper is structured as follows. The signal model and notation are presented in Sec. 2. In Sec. 3, the proposed method for solving permutations is introduced. In Sec. 4 experiments and results are presented. Finally, conclusions are summarized in Sec. 5.

2. Signal model and notation

Let us consider a time-invariant convolutive mixture of N independent speech sources, $s_j(n)$, with $j = 1, \ldots, N$, observed in an array of N sensors. In the absence of noise, each observation is given by

$$x_i(n) = \sum_{j=1}^{N} \sum_{k=-\infty}^{\infty} h_{ij}(k) s_j(n-k), \quad i = 1, \ldots, N, \tag{1}$$

where $h_{ij}(n)$ is the impulse response from jth source to ith sensor. If $X_i(f, t)$ and $S_i(f, t)$ are the STFTof $x_i(n)$ and $s_i(n)$, respectively, and $H_{ij}(f)$ is the Fourier transform of $h_{ij}(n)$, then

$$X_i(f,t) = \sum_{j=1}^{N} H_{ij}(f) S_j(f,t), \quad i = 1, \ldots, N. \tag{2}$$

By defining $\mathbf{X}(f,t) = [X_1(f,t), \ldots, X_N(f,t)]^T$, $\mathbf{S}(f,t) = [S_1(f,t), \ldots, S_N(f,t)]^T$, and $\mathbf{H}(f)$ so that $H_{ij}(f) = [\mathbf{H}(f)]_{ij}\ \forall i, j$, Eq. (2) can be rewritten as $\mathbf{X}(f,t) = \mathbf{H}(f)\mathbf{S}(f,t)$. The separation problem is then decoupled since there is an instantaneous mixture for each frequency bin. The separation matrices $\mathbf{B}(f)$ can be estimated independently for each frequency bin by any appropriated instantaneous separation method, thus providing the vector of outputs or estimated sources

$$\mathbf{Y}(f,t) = [Y_1(f,t), \ldots, Y_N(f,t)]^T = \mathbf{B}(f)\mathbf{X}(f,t). \tag{3}$$

Due to the scaling and ordering ambiguities in blind source separation problems, the recovered signals have an arbitrary (and, in general, different) permutation and scaling in each frequency bin. So the outputs vector is given by

$$\mathbf{Y}(f,t) \approx \mathbf{P}(\mathrm{f})\mathbf{D}(\mathrm{f})\mathbf{S}(\mathrm{f},\mathrm{t}), \tag{4}$$

where $\mathbf{P}$(f) is a permutation matrix and $\mathbf{D}(f)$ is a diagonal matrix of complex scalars. The scaling ambiguity, causing a filtering of the sources, is not a serious problem. Several methods can be used to reduce this effect; for example, the minimal distortion principle, where $\mathbf{B}(f)$ is replaced by $\mathrm{diag}\{\mathbf{B}(f)^{-1}\}\mathbf{B}(f)$. However, the permutation ambiguity is critical. Indeed, if the ordering of the estimated sources is not the same for every frequency bins, the transformation into time domain will be erroneous even when perfect separation is achieved. Therefore, to guarantee a constant ordering for all frequency bins, we have to estimate $\mathbf{P}(f)$ and to correct the permutations.

3. Proposed method

A key property of speech signals is their amplitude modulation correlation, i.e., the similarity or high correlation between all the frequency components of a speech signal when its STFT is transformed into a logarithmic scale.[3] This transformation is justified since perceived loudness is approximately logarithmic.[4] We define the normalized logarithmic magnitude of the jth output at the fth frequency bin as

$$Y_j^{dB}(f,t) = \frac{\log\left(|Y_j(f,t)|^2\right) - \left\langle \log\left(|Y_j(f,t)|^2\right)\right\rangle_t}{\sqrt{\left\langle \left|\log\left(|Y_j(f,t)|^2\right) - \left\langle \log\left(|Y_j(f,t)|^2\right)\right\rangle_t\right|^2\right\rangle_t}} \tag{5}$$

where $\langle\cdots\rangle t$ is the average over time. Note that due to this normalization, every base taken for the logarithm provides the same results. In these conditions, a good measure of similarity between two different frequency components is their cross-correlation,

$$C_{ij}(f_k, f_l) = \left\langle Y_i^{dB}(f_k,t)\, Y_j^{dB}(f_l,t)\right\rangle_t, \tag{6}$$

since from the entropy power inequality[5] the difference between logarithmic magnitude of different human voices can be approximated as Gaussian. Then, for each frequency bin f_k and for each output $Y_i^{dB}(f_k,t)$, a measure of its similarity with respect to all the frequency bins of another output, $Y_j^{dB}(f_l,t)$, is

$$\bar{C}_{ij}(f_k) = \sum_{\substack{l=1 \\ l\neq j}}^{nF} C_{ij}(f_k,f_l), \tag{7}$$

where n_F is the number of frequency bins in the STFT. Since two different components with the same frequency must belong to different outputs, the comparison between them does not make sense. For this reason, the frequency f_k should be excluded from the sum.

With the measure of similarity given by Eq. (7) we aim to determine the output corresponding to each frequency component. Following the optimal pairing principle,[6] the correct ordering for a frequency bin is given by the permutation that maximizes the sum of all the $\bar{C}_{ij}(f_k)$ corresponding to the permutation. With the aim of reordering the outputs at a frequency f_k, we can apply $N!$ possible permutation matrices. Let Π be one of these matrices, so that the new outputs vector at the frequency f_k is given by $\mathrm{Y}_\Pi(f,t) = \Pi \mathrm{Y}(f,t)$. We define the function

$$\rho_\Pi(f_k) = \sum_{(i,j)\in\Omega_\Pi} \bar{C}_{ij}(f_k), \tag{8}$$

where $\Omega_\Pi = \{(i,\ j)\colon [\Pi]_{i,j} = 1\}$ is the set of pairs of indexes involved in the permutation matrix Π. Then the correct order for the frequency bin f_k is given by

$$\Pi_o(f_k) = \arg\max_{\Pi} \rho_\Pi(f_k). \tag{9}$$

3.1 Improvement of the results

If we order the frequency components according to the rule expressed in Eq. (9), each frequency component will be assigned to the closest output. Generally, by repeating the process, all permutations are solved in a few iterations. However, in some cases, a few isolated permutations might still remain. This is due to the fact that the coherence between frequency components decreases as their distance (in frequency) increases. Therefore when all frequency bins have been assigned, we can solve these remaining permutations by defining a localized measure of similarity, emphasizing the neighboring frequencies. In this second stage, the measure of similarity given by Eq. (7) is then replaced by

$$\bar{C}_{ij}^{w}(f_k) = \frac{1}{n_F} \sum_{\substack{l=1 \\ l\neq j}}^{n_F} w(f_l,f_k) C_{ij}(f_k,f_l), \tag{10}$$

where $w(f_l,\ f_k)$ is a localized window, decreasing its value as $|f_l - f_k|$ increases. Now the function to be maximized, $\rho_\Pi(f_k)$, is substituted by

$$\rho_\Pi^{w}(f_k) = \sum_{i,j\in\Omega_\Pi} \bar{C}_{ij}^{w}(f_k). \tag{11}$$

This function should only be used at the second stage, when there are few and isolated permutations at the outputs, since the localized window is less robust to large or blockwise permutations.

3.2 Summary of the proposed algorithm

The algorithm consists of two stages: the first one compares each frequency component to every other frequency component of each output by means of Eq. (8); in the second one the neighboring frequencies are emphasized by using Eq. (11). The procedure is iterative, so that at each iteration, each frequency component is associated with the most similar output to this component. The algorithm can be summarized as follows.

Stage I: Comparison to all frequency components.

1. Calculate $\rho_{\mathbf{\Pi}}(f)$ for all possible permutations and for all frequency bins.
2. For $f_k = 1$: n_F.
   i) Find the reordering matrix $\Pi_0(f_k) = \arg\max_{\Pi} \rho_{\Pi}(f_k)$;
   ii) reorder the outputs as $\mathbf{Y}(f_k,t) \leftarrow \mathbf{\Pi}_0(f_k)\mathbf{Y}(f_k,\ t)$.
   EndFor
3. If $\mathbf{\Pi}_o(f_k) = \mathrm{I}\,\forall f_k$: GOTO Stage II;
   Else: GoTo step 1.

Stage II: Neighboring frequencies emphasized.

1. Calculate $\rho^w_{\mathbf{\Pi}}(f)$ for all possible permutations and for all frequency bins.
2. For $f_k = 1{:}n_F$.
   i) Find the reordering matrix $\mathbf{\Pi}^w_o(f_k) = \arg\max_{\Pi} \rho^w_{\Pi}(f_k)$;
   ii) reorder the outputs as $Y(f_k,t) \leftarrow \mathbf{\Pi}^w_o(f_k)\mathbf{Y}(f_k,t)$.
   EndFor
3. If $\mathbf{\Pi}^w_o(f_k) = I\ \forall f_k$: END of the ALGORITHM;
   Else: GoTo step 1.

3.3 Computational cost

The computational cost of the proposed method is determined by the computation of $\bar{C}_{ij}(f_k)$, and is of order $O\left(N^2 n_F^2 T\right)$, where T is the number of time bins. The computation of $\rho_{\mathbf{\Pi}}(f_k)$ has a cost of order O ($N!N\ n_F$). Since the number of frequency bins is much greater than the number of sources ($n_F \gg N$), the first computation is the limiting one. Nevertheless, it should be emphasized that the global complexity of the complete separation process in frequency domain is dominated by the computational complexity of the separation algorithms.

4. Simulations

We performed two set of experiments. The first one illustrates the performance of the proposed algorithm in a situation of perfect separation (i.e., when, at each frequency bin, the sources are completely separated at the outputs, but randomly permuted). The second one tests the performance of the proposed method with a live recording.

4.1 Performance in a perfect separation situation

In order to illustrate the performance of the proposed method after a perfect separation, we applied permutation matrices, $\mathbf{P}(f)$, to a set of speech sources. We did several experiments, for $N = 2,\ldots,7$ sources, running the proposed algorithm in order to recover the original spectrograms. For each experiment, we made 30 simulations, randomly selecting the permutation matrix for each frequency. We used male and female sources, with a duration of 5 s (sampled at 10 KHz), randomly chosen from the database http://www.imm.dtu.dk/pubdb/p.php?4400 of 12 individual recordings. The STFT were computed using Hanning windows of length 1024 samples, FFT of 2048 points, and 90% overlap, giving a number of 479 time bins, which are used for the temporal average. For the second stage of the proposed method we used as localized window in frequency a Hamming window of 1024 samples that was truncated when necessary. The results showed that, for all simulations, the proposed method completely reordered

Table 1. Results for a perfect separation situation with $N=2$ sources. 30 simulations were made by applying a randomly selected permutation matrix to each frequency bin. The proposed algorithm was compared with two other (Refs. 3, 8). The averaged number of remaining permutations (errors) per simulation and the number of simulations for which there is one remaining permutation at least are shown.

	Average number of errors per simulation	Number of simulations with errors
Proposed	1.8	7
Pham-Servière (Ref. 3)	5.3	15
Rahbar-Reilly (Ref. 8)	121.4	23

the frequency components, except for certain very low frequencies (lower than 50 Hz) that remained permuted for some pairs of speakers. This is due to the fact that for very low frequencies the sources do not always satisfy the property of spectral coherence. However, this does not affect the quality of the recovered sources.[7] For the 2×2 case we show a comparison with two other methods[3,8] (see results in Table 1). The proposed method exhibited the smallest number of unsolved permutations, both on average (about three times smaller than the method proposed in Ref. 3) and in each run. Also, it fails less (we consider a fail when there is at least one unsolved permutation). So we can conclude that this method outperforms the others. We also made 30 simulations for each of the cases with $N=3,\ldots,7$ sources, obtaining very good results, with a maximum for the average number of unsolved permutations per simulation of 24.1 (7×7 case) and a minimum of 2.6 (3×3 case), always for very low frequencies. Results are shown in Fig. 1.

4.2 Performance for live recording

With the aim of illustrating the behavior of the proposed algorithm in a real environment we applied it to a live recording of 10 s, with a sample rate of 16 kHz, provided in http://sisec2010.wiki.irisa.fr/tiki-index.php. Three speech sources were played and then recovered by three omnidirectional microphones in a room of 3.55 m of width, 4.45 of length, and 2.5 m of height. The distance between sources and microphones was around 1 m, and the maximum distance between two microphones was 5.7 cm (corresponding to a delay of about 1/6 ms, greater than the sample period). The STFT was computed by using a window length of 2048 samples, an FFT of 4096 points, and an overlap of 95%, resulting in 1543 time bins. The original sources were estimated from the observations by means of the THINICA algorithm using the initialization procedure proposed by us in Ref. 7 followed by the proposed method for permutations

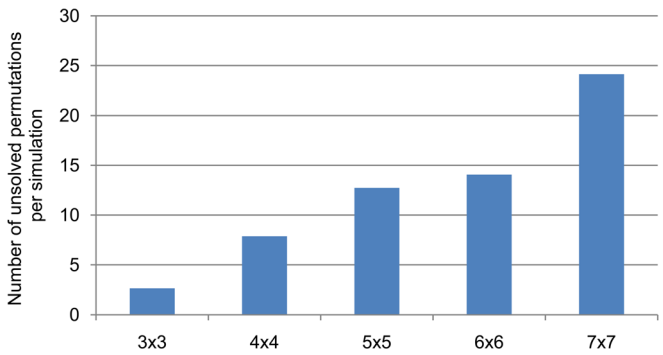

Fig. 1. (Color online) Performance of the proposed algorithm in a situation of perfect separation when the number of sources are $N=3,\ldots,7$. The number of remaining permutations per simulation are represented for different cases.

correction, which used 1.1% of the global process duration. The quality of the estimated sources was measured in terms of source to interferences ratio (SIR) by E. Vincent, since the original sources are not public; an average SIR of 10.1 dB was obtained. The observations can be listened to in Mm. 1, Mm. 2, and Mm. 3, while the estimated sources can be listened to in Mm. 4, Mm. 5, and Mm. 6.

Mm. 1. Observation 1. Live recording sampled at 16 kHz. This is a file of type "wav" (312 kb)

Mm. 2. Observation 2. Live recording sampled at 16 kHz. This is a file of type "wav" (312 kb).

Mm. 3. Observation 3. Live recording sampled at 16 kHz. This is a file of type "wav" (312 kb).

Mm. 4. Mm. 4: Estimated source 1. This is a file of type "wav" (312 kb).

Mm. 5. Mm. 5: Estimated source 2. This is a file of type "wav" (312 kb).

Mm. 6. Mm. 6: Estimated source 3. This is a file of type "wav" (312 kb).

5. Conclusions

We proposed a method for solving the permutation problem that arises in blind $N \times N$ separation of convolved speech signals when working in the time-frequency domain. The proposed method combines the assumption of spectral coherence for the speech sources and an optimal pairing scheme. We defined for each frequency bin a measure of coherence based on the amplitude modulation correlation property and derived a method which consisted of two stages: the first one considers all the frequency bins, whereas the second one emphasizes the neighboring frequencies. We illustrated the good performance of the proposed method in terms of the SIR by means of a set of experiments for perfect separation situations and for live recordings.

Acknowledgments

This work was supported by Ministry of Science and Innovation of Spain through Project No. TEC2011-23559. We thank Emmanuel Vincent's collaboration for the evaluation of the results.

References and links

[1]M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in *Springer Handbook of Speech Processing* (Springer, Berlin, 2008).
[2]J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation*, (June, 2000), pp. 215–220.
[3]D. T. Pham, C. Serviére, and H. Boumaraf, "Blind separation of convolutive audio mixtures using nonstationarity," in *Proceedings of ICA 2003 Conference*, Nara, Japan (April, 2003).
[4]L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition* (Prentice Hall, Englewood Cliffs, NJ, 1993).
[5]S. Curces, A. Cichocki, and S.-i. Amari, "From blind signal extraction to blind instantaneous signal separation," IEEE Trans. Neural Networks **15**(4), 859–873 (2004).
[6]P. Tichavsky and Z. Koldovsky, "Optimal paring of signal components separated by blind techniques," IEEE Signal Process. Lett. **11**(2), 119–122 (2004).
[7]A. Sarmiento, I. Durán-Díaz, and S. Cruces, "Initialization method for speech separation algorithms that work in the time frequency domain," J. Acoust. Soc. Am. **127**(4), 121–126 (2010).
[8]K. Rahbar and J. P. Reilly, "A frequency domain method for blind source separation of con- volutive audio mixtures," IEEE Trans. Speech Audio Process. **13**(5), 832–844 (2005).