A NOVEL ROBUST SOLUTION TO THE PERMUTATION PROBLEM BASED ON A JOINT MULTIPLE TDOA ESTIMATION

F. Nesta

[1]Fondazione Bruno Kessler - irst, [2]UNITN
Trento (TN), Italy
Email: nesta@fbk.eu

M. Omologo, P. Svaizer

Fondazione Bruno Kessler - irst
Trento (TN), Italy
Email: {omologo, svaizer}@fbk.eu

ABSTRACT

This paper proposes a new robust method to perform a multiple TDOA estimation in order to solve the permutation problem in frequency-domain Blind Source Separation. According to the acoustic propagation model, in frequency-domain, each separation matrix can be represented with a set of states associated with each source. A novel transform of the states is introduced which is independent of the aliasing and of the permutations and is able to perform a joint estimation of multiple TDOAs. We show that such a transform generalizes the GCC-PHAT for multiple sources and at the same time generates envelopes with clear peaks corresponding to the maximum likelihood TDOAs. By means of the propagation model, the permutation problem is solved using the estimated TDOAs. Experimental results show that the proposed approach allows one to separate two speakers, using very short utterances (0.5-1s), in highly reverberant environment ($T_{60} = 700ms$) even with widely-spaced microphones.

Index Terms— permutation problem, blind source separation (BSS), TDOA estimation, independent component analysis (ICA), speech enhancement, multiple sources localization

1. INTRODUCTION

Among different approaches of source separation, the multichannel frequency-domain is maybe the most investigated for its fast convergence property and reasonable computational complexity. During the last years a huge number of methods have been proposed but separation in real-life is still an open problem. Frequency-domain approaches can directly exploit the advantages of Independent Component Analysis and long demixing filters can be estimated in a reasonable time. However, unlike time-domain approaches, the ambiguity of permutation must be solved and is still an open problem. Among the most promising ones a robust way to solve the permutation problem is to estimate the propagation model parameters of the sources [1]. This approach is suitable for the case of separation of short utterances since permutations are solved by using only the estimated demixing matrices. However reverberation and spatial aliasing makes the estimation quite sensitive to errors.

In this work we present a new method able to estimate multiple TDOAs, which is robust both to spatial aliasing and to reverberation even for high T_{60}. A novel transform of the state space associated with the separation matrices is formulated. This transform, which will be called State Coherence Transform (SCT), jointly exploits states associated with all the sources and thus is invariant to permutations. By using the SCT, without requiring any knowledge of the maximum distance between microphones, we generate envelopes with clear peaks corresponding to the TDOAs of the sources. In the following section a physical interpretation of the separation matrix $\mathbf{W}(f)$ is recalled. The novel SCT transform is presented and its permutation-invariance property is then explained. Furthermore it is demonstrated that the SCT is a generalization of the GCC-PHAT [2] for the multidimensional case. Finally experimental results, reported in section 5, confirm that the proposed method can solve the permutation problem under challenging conditions allowing the separation of short utterances in highly reverberant environments.

2. PHYSICAL INTERPRETATION OF W(*f*)

To simplify the understanding of the proposed approach we first recall a simple physical interpretation of the separation matrices. Assuming the sources to be under free-field conditions, signals acquired through the microphones are delayed and scaled versions of the original ones. In the frequency-domain the observed components $\mathbf{y}(f)$ can be modeled as:

$$\mathbf{y}(f) = \mathbf{H}(f)\mathbf{x}(f) \tag{1}$$

where $\mathbf{H}(f)$ is a complex-valued matrix and $\mathbf{x}(f)$ are the frequency components of the original signals. In the frequency-domain BSS each frequency component is separated by estimating a separation matrix $\mathbf{W}(f)$. In our earlier work [3] we showed that, in the case of two microphones, the ratios of the rows of $\mathbf{W}(f)$ are scaling invariant and can be viewed as observations of the propagation models associated with each source. In general the observations of the propagation models can be obtained by the ratio of the columns of $\mathbf{W}(f)^{-1}$, which can be viewed as an estimate of $\mathbf{H}(f)$, up to a permutation and scaling ambiguity:

$$r_k(f) = \frac{w_{ak}^{-1}}{w_{bk}^{-1}} = |r_k(f)|e^{-j2\pi f \Delta t_k} \tag{2}$$

where k is the index of the source, w_{ik}^{-1} are the elements of the demixing matrix $\mathbf{W}(f)^{-1}$, and Δt_k is the observed TDOA for the $k-th$ source with respect to the chosen microphone pair $a-b$. Each ratio depends on the frequency and on the TDOA and thus can be considered as a state associated to each source. If we assume that the permutation problem is solved, using states associated with different frequencies, it is possible to estimate the TDOA for each source. This evidence was exploited in [1] to group the frequency components using an effective estimation of the propagation model parameters. Such a method estimates the parameters for the frequencies where the spatial aliasing does not occur, grouping the states belonging to the same source, according to a procedure similar to k-means. After that, it recursively solves the permutations for the higher frequencies refining at each step the estimation of the model parameters. Although this method is effective in normal situations it lacks reliability if very short utterances are analyzed with high reverberation. In fact, with the recursive structure of the approach, if there are some frequency bands for which the ICA solution is not reliable, the model parameters are incorrectly refined and this generates wrong permutation decisions. Thus a more reliable estimation is needed which takes into account all the observed states at the same time without any recursive schema.

Before presenting the formulation of the SCT we first recall the general optimization rule that can be used to estimate the model when the permutation does not occur. Under ideal conditions the frequency-domain model of the inter-microphone delay for a given source can be represented by:

$$c(f,\tau) = e^{-j2\pi f\tau} \tag{3}$$

where τ is the relative time difference of arrival between the two microphones. Thus, assuming that the permutations have been solved, an estimation of the TDOA associated with each source can be performed by minimizing the following quantity:

$$\overline{\tau}_k = \underset{\tau}{\operatorname{argmin}} \int ||c(f,\tau) - \overline{r}_k(f)||df \tag{4}$$

where $\overline{r}_k(f)$ are the normalized states computed as:

$$\overline{r}_k(f) = \frac{r_k(f)}{||r_k(f)||} \tag{5}$$

Once the TDOAs (τ_k) and thus the model $c(f,\tau_k)$ are known, the permutation alignment is a trivial step. Each frequency component is aligned according to the permutation that minimizes the following quantity:

$$\overline{\Pi}(f) = \underset{\Pi}{\operatorname{argmin}} \sum_{k=1}^{N} ||c(f,\tau_k) - \overline{r}_{\Pi_{(k)}}(f)|| \tag{6}$$

where Π is a permutation of the matrix $\mathbf{W}(f)$ and N the number of the sources.

If the permutation problem is not solved, the state $\overline{r}_k(f)$ does not always belong to the same source and the estimation cannot be performed directly with (4). However we will show in the following section that, regardless of permutations, it is possible to formulate a transform that takes into account all the observed states and exhibits maxima located exactly at the TDOAs related to all the observed sources.

3. STATE COHERENCE TRANSFORM

The State Coherence Transform is formulated as follows:

$$SCT(\tau) = \int \sum_{k=1}^{N} \left[1 - g\left(\frac{||c(f,\tau) - \overline{r}_k(f)||}{2}\right)\right] df \tag{7}$$

where N is the number of rows of matrix $\mathbf{W}(f)$ and $g(\cdot)$ is a function of the euclidean distance. To better understand the behavior of the above transform, we first define an approximated SCT (aSCT) considering $g(\cdot)$ equal to:

$$g(x) = x \tag{8}$$

Thus the original formula (7) can be simplified to:

$$aSCT(\tau) = \int \sum_{k=1}^{N} \left(1 - \frac{||c(f,\tau) - \overline{r}_k(f)||}{2}\right) df \tag{9}$$

The aSCT will be maximized for values of τ that for each frequency minimize the sum of the distances between the vector $c(f,\tau)$ and all the observed $\overline{r}_k(f)$:

$$d(f,\tau) = \sum_{k=1}^{N} ||c(f,\tau) - \overline{r}_k(f)|| \tag{10}$$

It is simple to demonstrate that for the case of $N = 2$ (two microphones), for a given frequency, the above distance $d(f,\tau)$ can be minimized with the values τ_l that solve each equation:

$$||c(f,\tau) - \overline{r}_l(f)|| = 0, \quad \forall l = 1..2 \tag{11}$$

If $N = 2$ formula (10) becomes:

$$d(f,\tau) = ||c(f,\tau) - \overline{r}_1(f)|| + ||c(f,\tau) - \overline{r}_2(f)|| \tag{12}$$

For each of the values τ_1 or τ_2, which solve equation (11), the distance $d(f,\tau)$ reduces to $||\overline{r}_1(f) - \overline{r}_2(f)||$ since $c(f,\tau_l) = \overline{r}_l(f)$ and thus one of the terms of (12) would be zero.

Furthermore, all the vectors have unit norm and thus, with this geometrical constraint, for each τ there must be:

$$||\overline{r}_1(f) - \overline{r}_2(f)|| \leq d(f,\tau) \tag{13}$$

Thus, values of τ for which the distance $d(f,\tau)$ is minimized for most of the frequencies f, maximize the approximated SCT in (9) and generate clear peaks in its corresponding envelope. Such values of τ will represent the maximum-likelihood TDOA of each source, according to the model in (3).

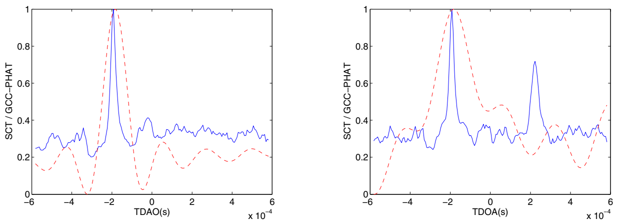
For the case of $N > 2$ we cannot define an inequality like in (13). For this reason the aSCT would lead to errors, since it would be maximized even for τ values that are a linear combination of the TDOAs of the sources. To solve this problem we need a function $g(\cdot)$ that is able to give more emphasis to the states $\overline{r}_k(f)$ that are closer to the model $c(f,\tau)$ while it neglects the others.

We empirically found that a sigmoid function, like the $\tanh(\cdot)$, gives a good non-linear transformation of the euclidean distance. We can thus define $g(\cdot)$ as:

$$g(x) = \tanh(\alpha \cdot x) \tag{14}$$

where α is a positive real-valued shape factor which modifies the inter-source TDOA resolution.

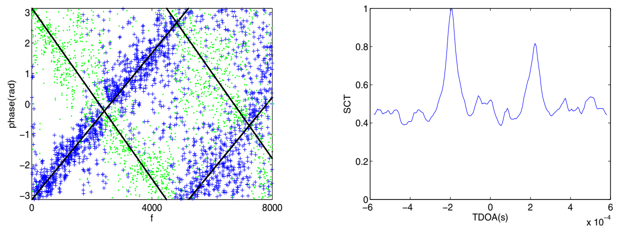
To show the effectiveness of the proposed method we consider a real situation and we apply both the SCT and the GCC-PHAT (also known as CSP [4]) in order to perform a TDOA estimation. Two speakers, placed 2 meters away from



(a) One active source

(b) Two active sources

Fig. 1. Comparison between the SCT (solid line) and GCC-PHAT (dashed line) envelopes.



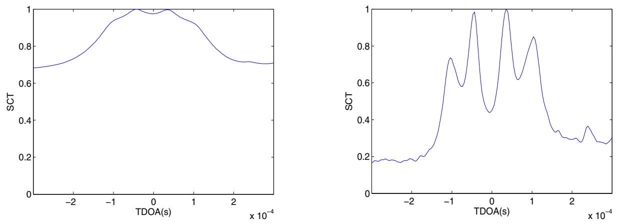
(a) states $\overline{r}_k(f)$ and estimated models $c(f, \tau_k)$ (black lines)

(b) SCT profile of the state $\overline{r}_k(f)$

Fig. 2. Plot of the states $\overline{r}_k(f)$ and its SCT transform for a real test.

the array and with an angular distance of about 40°, have been recorded using two microphones, spaced at 20 cm, in a room with a T_{60} of about $700ms$. For the SCT the demixing matrices, and thus the states $\overline{r}_k(f)$, are obtained using the algorithm presented in [3] using just one second of data and an FFT window size of 4096 taps. The GCC-PHAT is computed by framing and averaging FFT windows of 4096 taps, overlapped of 3840 taps. The values of τ for which the SCT is computed are chosen in order to have a theoretical spatial resolution of 1°. The GCC-PHAT is interpolated with a factor of 10 to have a comparable resolution. In both cases signals are sampled at Fs=$16kHz$ and the envelopes are normalized to 1.

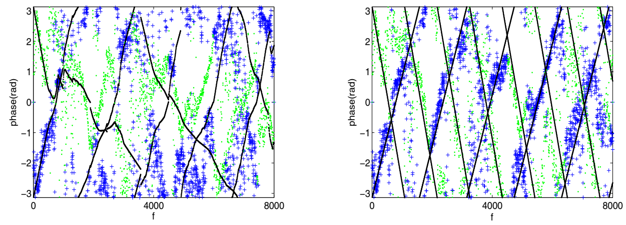
In figure 1 we compare the envelopes obtained by using the SCT (solid line) and the GCC-PHAT (dashed line) when one and two sources are active, respectively. In the first case both envelopes exhibit a peak located almost at the same position and this confirms that the SCT leads to similar results as the GCC-PHAT does when just one source is active. In the second case the GCC-PHAT is not able to detect the secondary peak whilst the SCT exhibits clear peaks corresponding to the TDOAs of the two sources (expected to be about -0.2 and +0.2 ms). In figure 2 we can observe the plot of the phase of the states $\overline{r}_k(f)$ and the profile of the SCT transform. The black lines represent the models $c(f, \tau_k)$ estimated for each source where τ_k are the TDOA selected by means of the peaks of the SCT envelope. It is worth noting that such lines accurately approximate the states $\overline{r}_k(f)$ and thus, as we expected, the values τ_k are reliable estimations of the TDOA of each source and can be effectively used to solve the permutation problem according to the rule in (6). In figure 3 the SCT envelopes for the case of $N = 4$ sources are plotted. Two different values of α have been chosen in order to demonstrate that the non-linear mapping can effectively increase the resolution between the sources. Finally, figure 4 shows that a recursive re-estimation like in [1] is not always able to converge to a correct TDOA estimate whilst the SCT, exploiting all the frequencies at the same time, provides a stable result.



(a) $\alpha = 1$

(b) $\alpha = 10$

Fig. 3. SCT envelopes computed for the case of $N = 4$ sources, using different values of α.



(a) $c(f, \tau_k)$ re-estimated with a recursive approach similiar to [1]

(b) $c(f, \tau_k)$ estimated by means of the SCT peaks

Fig. 4. Phase-frequency plot of the states $\overline{r}_k(f)$ and of the estimated models $c(f, \tau_k)$ (black lines).

4. CONNECTIONS BETWEEN SCT AND GCC-PHAT

When just one state observation is available, the aSCT is equivalent to the GCC-PHAT. To simplify the notation the aSCT can be modified using the squared distance between the model $c(f, \tau)$ and the state $r_k(f)$ without changing the meaning of the obtained envelopes. Thus for the case of $N = 1$, (9) becomes:

$$\overline{aSCT}(\tau) = \int \left(1 - \frac{||c(f,\tau) - r_1(f)||^2}{2}\right) df \qquad (15)$$

By a simple mathematical manipulation it can be written as:

$$\begin{aligned}\overline{aSCT}(\tau) &= \int \left(1 - \frac{[c(f,\tau) - r_1(f)][c(f,\tau) - r_1(f)]^*}{2}\right) df \\ &= \int (Re[c(f,\tau)^* r_1(f)])\, df \qquad (16)\end{aligned}$$

Using the signals observed with a pair of microphones, the GCC-PHAT is computed as:

$$GCC - PHAT(\tau) = \int \left(\frac{x_1(f)x_2(f)^*}{||x_1(f)x_2(f)^*||}\right) e^{j2\pi f\tau} df \qquad (17)$$

where $x_1(f)$ and $x_2(f)$ are the Fourier transforms of the signals recorded by the first and the second microphone, respectively. Under ideal conditions each microphone observes a delayed version of the original acoustic wave according to the position of the source. So for each frequency the product $x_1(f)x_2(f)^*$ can be rewritten as:

$$\begin{aligned}(x_1(f)x_2(f)^*) &= |x_1|e^{-j2\pi f T_1}|x_2|e^{j2\pi f T_2} \\ &= |x_1|e^{-j2\pi f T_1}|x_2|e^{[j2\pi f(T_1+\delta\tau)]} \\ &= |x_1||x_2|e^{j2\pi f\delta\tau} \qquad (18)\end{aligned}$$

where $\delta\tau$ is the relative TDOA and T_1 and T_2 are the time of arrivals (TOA) of the direct wave front recorded by the microphones 1 and 2, respectively. Thus (17) can be simplified to:

$$GCC - PHAT(\tau) = \int e^{j2\pi f\delta\tau} e^{j2\pi f\tau} df \qquad (19)$$

The first exponential has the same meaning as the observed state $\overline{r}_1(f)$ since it represents the sound propagation model from the source to the two microphones. It is worth noting that PHAT normalization has a similar effect as the state normalization in (5). The second exponential instead is exactly the conjugate of $c(f,\tau)$. So without losing any generality (17) can be rewritten as:

$$\int c(f,\tau)^* r_1(f) df \qquad (20)$$

The only difference between (20) and (16) is the $Re[\cdot]$ operator and the two formulas in our problem are equivalent. In fact the integral in (20) is maximized for values that minimize the phase difference between $c(f,\tau)$ and $r_1(f)$. For such values the imaginary part goes to zero and thus it is equivalent to consider just the real part as in (16).

5. EXPERIMENTAL RESULTS

In order to demonstrate the effectiveness of the proposed method the SCT transform was applied to the recursive FD-BSS, presented in [3]. That algorithm is able to intrinsically solve the permutations when spatial aliasing does not occur. In this work however we use widely-spaced array and the permutation problem is solved by a posterior alignment according to the estimated TDOAs and the optimization in formula (6). The algorithm has been implemented both in Matlab and in C++ and is able to work in real-time with two sources. The ICA step is realized by means of the Scaled Infomax proposed in [5]. The algorithm has been applied to separate two sources with an angular distance of 45°, placed about 1.1 meter away from the array. The distance between the microphones was 25 cm and the test room has a $T_{60} = 700ms$. The sampling frequency is Fs=$16kHz$ and the length of the demixing filters has been chosen between 4096 and 8192 taps according to the training window size. Performance is evaluated according to the Signal-to-Interference Ratio (SIR). SIR values are computed using the whole signals (length of about 9s) but the demixing filters are computed using different training window sizes. The reported SIR is the average value over all the separated sources. Audio samples of the original and separated signals are available at [6].

As shown in table 1, high separation performance has been obtained in highly reverberant environment and with very short utterances (6 dB with just 500 ms of data and 7dB with 1 second of data). A comparison with the time-domain Parra's method [7] is also provided. Performance confirms that using frequency-domain methods, if the permutation problem is effectively solved, good results can be obtained even under challenging conditions.

training window size (s)	0.5	1	2	4	9
Proposed (dB)	6.01	7.02	7.94	10.9	11.5
Parra's (dB)	1.41	2.39	2.69	2.8	2.95

Table 1. SIR performance comparison

6. CONCLUSIONS

In this paper a robust solution to the permutation problem was presented, which uses a joint-multiple TDOA estimation of the sources. The method performs a non-linear transformation of the state-space based on its frequency coherence. Experimental results show that the approach can be successfully applied to solve the permutation problem even under challenging conditions. Furthermore, since permutations are solved without using any information of the non-stationarity of the signals, the method is stable even when using short utterances and long FFT windows to cope with high T_{60}.

7. REFERENCES

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

[2] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1976, vol. 24, pp. 320–327.

[3] F. Nesta, M. Omologo, and P. Svaizer, "Separating short signals in highly reverberant environment by a recursive frequency-domain bss," in *HSCMA*, Trento, Italy, May 2008.

[4] M. Omologo and P. Svaizer, "Acoustic event localization using crosspower spectrum phase based technique," in *Proc. ICASSP '94*, Adelaide, Austrailia, 1994, pp. II–273–II–276.

[5] S.C. Douglas and M. Gupta, "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation," in *ICASSP*, Apr. 2007, vol. II, pp. 637–640.

[6] http://shine.fbk.eu/people/nesta.

[7] L. Parra and C. Spence, "Convolutive blind source separation of non-stationary sources," *IEEE Trans. on Speech and Audio Processing*, pp. 320–327, May 2000.