ROBUST TWO-CHANNEL TDOA ESTIMATION FOR MULTIPLE SPEAKER LOCALIZATION BY USING RECURSIVE ICA AND A STATE COHERENCE TRANSFORM

F. Nesta

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

P. Svaizer, M. Omologo

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

ABSTRACT

A novel method is presented for a robust two channel multiple Time Difference of Arrival (TDOA) estimation for multi-speaker localization which can provide satisfactory performance even in highly reverberant environment. The method is based on a recursive frequency-domain Independent Component Analysis (ICA) and on a novel State Coherence Transform (SCT). Exploiting the phase coherence of the demixing matrices obtained in the ICA stage the SCT is able to generate envelopes with clear peaks in the corresponding maximum-likelihood TDOAs. The SCT envelopes are computed independently in each time-block and accurate multiple TDOAs are estimated by means of a time-frequency sparse representation of the sources. The method has been applied to real data obtained by recording many sources in a room with a reverberation time of 700ms. Experimental results show that an accurate localization of 7 closely-spaced sources is possibile given only few seconds of data even in the case of low SNR. Experiments also show the advantage of using the proposed solution rather than the well-known GCC-PHAT.

Index Terms— blind source separation (BSS), TDOA estimation, independent component analysis (ICA), multiple speaker localization

1. INTRODUCTION

Multiple speaker localization is a difficult problem which gives rise to high interest in the field of acoustic signal processing and audiovisual information fusion, in particular for meeting scenarios. A wide literature is available with this regard. A high reverberation time, the presence of strong enviromental noise, and spatial ambiguity make the localization task harder, especially when just two microphones are used. In the last years multiple TDOA estimation was also addressed by the BSS community since knowledge of the TDOAs is essential for underdetermined sound source separation.
Recent works show that the frequency-domain BSS is strictly connected with the wave propagation from the sources [1]. Our earlier work [2] showed that a joint TDOA estimation can be performed for all the sources by using the demixing matrices, estimated for each frequency by an Independent Component Analysis algorithm. Such estimation is accomplished by a proper State Coherence Transform (SCT) of the state space associated with the demixing matrices. This is invariant to the permutation problem and it is insensitive to the spatial aliasing caused by phase-wrapping.
In this work we extend the SCT to a cumulative SCT (cSCT) based on a sparse-dominance assumption of the sources, which is able to estimate TDOAs related to many sources by using only two microphones. The next section recalls the physical interpretation of the ICA when applied in frequency-domain since it is the starting point of the proposed method.

2. FREQUENCY-DOMAIN ICA AND ITS PHYSICAL INTERPRETATION

A straightforward interpretation of the frequency-domain ICA is given when applied to two channel mixtures of two observed sources. The signals observed by the microphones can be modeled using a time-frequency representation where each component is evaluated by a short-time Fourier analysis. For each frequency a time observation can be considered as a linear combination of the time-frequency components associated to the original source signals. In matrix notation one can write:

$$\mathbf{y}(k,\tau) = \mathbf{H}(k)\mathbf{x}(k,\tau) \tag{1}$$

where $\mathbf{y}(k,\tau)$ are the observed mixtures, $\mathbf{x}(k,\tau)$ are the original signals, τ is the time instant at which each frequency is evaluated according to the time-frame shifting, k is the frequency bin index and $\mathbf{H}(k)$ is a mixing matrix. Thus, by applying a complex-valued ICA to the time-series of each frequency, the original components can be retrieved by computing a demixing matrix $\mathbf{W}(k)$ which is an estimate of the matrix $\mathbf{H}(k)^{-1}$ up to scaling and permutation ambiguities:

$$\overline{\mathbf{x}}(k,\tau) = \Lambda(k)\mathbf{P}(k)\mathbf{W}(k)\mathbf{y}(k,\tau) \tag{2}$$

where $\Lambda(k)$ and $\mathbf{P}(k)$ are a complex-valued scaling matrix and a permutation matrix, respectively.

In anechoic ideal environment the mixing matrix could be modeled as:

$$\mathbf{H}(k) = \begin{pmatrix} |h_{11}(k)|e^{-j\varphi_{11}(k)} & |h_{12}(k)|e^{-j\varphi_{12}(k)} \\ |h_{21}(k)|e^{-j\varphi_{21}(k)} & |h_{22}(k)|e^{-j\varphi_{22}(k)} \end{pmatrix} \tag{3}$$

$$\varphi_{iq}(k) = 2\pi f_k T_{iq} \tag{4}$$

where T_{iq} is the propagation time delay from the *q-th* source to the *i-th* microphone and f_k is the true frequency associated to the *k-th* frequency bin. Thus, if the reverberation is neglected, the phase term φ_{iq} is expected to vary linearly according to the frequency and such a linearity can also be found in the estimated separation matrices $\mathbf{W}(k)$. In fact our earlier work [3] has shown that the ratios between the elements of

each row of $\mathbf{W}(k)$ are scaling invariant and can be considered as observations of the ideal propagation model of the acoustic wave related to each source:

$$r_1(k) = \frac{|h_{12}|}{|h_{22}|} e^{-j2\pi f_k \Delta t_1}, \;\; r_2(k) = \frac{|h_{11}|}{|h_{21}|} e^{-j2\pi f_k \Delta t_2} \quad (5)$$

where Δt_1 and Δt_2 are the true TDOAs of the sources. Each ratio depends on the frequency and on the TDOA and thus can be considered as a state associated to each source. Assuming the sources to be in far-field conditions, the propagation model of a source yelding a TDOA of τ can be represented as:

$$c(k,\tau) = e^{-j2\pi f_k \tau} \quad (6)$$

In real conditions the acoustic propagation is distorted by the reverberation effects and the ratios $r_i(k)$ are noisy observations of the ideal propagation models. Each TDOA can be effectively estimated by minimizing the euclidean distance between the ideal model and the normalized ratios:

$$\overline{\Delta t}_i = \underset{\tau}{\operatorname{argmin}} \sum_k ||c(k,\tau) - \overline{r}_i(k)|| \quad (7)$$

where $\overline{\Delta t}_i$ is the TDOA estimated for the *i-th* source and the states are normalized as follows:

$$\overline{r}_i(k) = \frac{r_i(k)}{||r_i(k)||} \quad (8)$$

However for the permutation ambiguity we do not know which states belong to a particular source and the TDOA estimation cannot be directly performed with the minimization in (7). In [4] we showed that a multiple TDOA estimation can be performed by a proper transform which jointly uses the states associated to all the sources and thus it is invariant to the permutations. Such a transform was referred to as State Coherence Transform and was formulated as follows:

$$SCT(\tau) = \sum_k \sum_{i=1}^{N} \left[1 - g\left(\frac{||c(k,\tau) - \overline{r}_i(k)||}{2} \right) \right] \quad (9)$$

where N is the number of observed states for each frequency and $g(\cdot)$ is a function of the euclidean distance. In [4] it has been shown that, in the case $N = 2$ by choosing $g(x) = x$ a mathematical constraint holds for which the SCT will always be maximized for τ equal to the maximum-likelihood TDOA of each source.

3. CUMULATIVE SCT FOR MULTIPLE TDOA ESTIMATION

The SCT is theoretically able to estimate a number of TDOAs at least equal to the number of microphones. For the case of two channels, for the *k-th* frequency bin, ICA estimates two ratios $r_1(k)$ and $r_2(k)$ which are expected to represent the propagation of two sources. However when the number of the sources is greater than the number of the microphones we can assume a sparse dominance of the sources in the time-frequency domain. This assumption holds, for example, in the case of multiple speakers emitting sound with comparable powers. Hence, for each frequency ICA estimates a demixing matrix which represents an observation of the propagation models associated to the two dominant sources at a given instant. By computing ICA in different time-frequency blocks we expect to observe states which represent the propagation models of all the active sources. Then the coherence of such states can be globally evaluated by using a cumulative SCT (cSCT), which is formulated as follows:

$$cSCT(\tau) = \sum_b \sum_k \sum_{i=1}^{N} \left[1 - g\left(\frac{||c(k,\tau) - \overline{r}_i^b(k)||}{2} \right) \right] \quad (10)$$

where $\overline{r}_i^b(k)$ is the normalized state obtained for the $k-th$ frequency bin in the time block b. The peak positions in the cSCT envelope reveal the TDOAs of the active sources
It is worth noting that the SCT can also be derived if ICA is applied to a number of microphones larger than two. In this work we focused on the two channel estimation problem since we observed that two microphones are sufficient to enable the TDOA estimation for a considerable number of sources. However, it is clear that as the number of the sources is increased, the sparsity of the signals in time-frequency domain decreases. Hence the observed ratios $\overline{r}_i^b(k)$ would be more accurate if a larger number of microphones were used in the ICA stage.
The theoretical formulation of (10) requires the estimation of the ratios $\overline{r}_i^b(k)$ to be defined for different time-blocks. However since ICA needs to be applied to relatively long signals, a trade-off between time resolution and ICA accuracy is needed. In [5] we showed that a recursive approach across the frequency can be exploited to increase the ICA accuracy when short signals are observed. In this work such a method allowed us to estimate the ratios $\overline{r}_i^b(k)$ even using time-blocks of less than $300ms$. The next section summarizes the main steps of the recursive ICA approach which plays an important role in the effectiveness of the cumulative SCT analysis.

4. RECURSIVE ICA

In [5] and [3] we proposed a new recursive approach to improve ICA when used for a frequency-domain blind source separation. In this work we are not interested in the source separation but only in the TDOA estimation by means of the separation matrices derived by the ICA step. The proposed strategy consists in performing the ICA recursively, e.g. from the highest to the lowest frequency. At each frequency a matrix $\mathbf{W}_{smooth}(k)$ is estimated as a smooth extension from the noisy matrices $\mathbf{W}(k)$ observed at previous frequencies. The resulting matrix is used to initialize the matrix $\mathbf{W}_0(k)$ for the ICA of the adjacent frequency. The matrix $\mathbf{W}_{smooth}(k)$ is estimated by filtering its determinant value. A simple and computationally inexpensive ε- Normalized Least Mean Square ($\varepsilon - NLMS$) predictor is used. In order to reduce the steady-state error, the LMS filter has been implemented with a variable ε approach as proposed in [6]. The full description of the proposed filtering procedure is presented as follows. For a given frequency bin k, in order to remove the scaling ambiguity, the observed matrix $\mathbf{W}(k)$ is normalized as:

$$\overline{\mathbf{W}}(k) = \mathbf{C}(k)\mathbf{W}(k) \quad (11)$$

where $\mathbf{C}(k)$ is computed as:

$$\mathbf{C}(k) = \begin{pmatrix} c_1(k) & 0 \\ 0 & c_2(k) \end{pmatrix}, \; c_i(k) = \frac{e^{-j \cdot arg(w_{ii}(k))}}{\sum_{n=1}^{N} |w_{in}(k)|} \quad (12)$$

The smooth matrices are computed, with k starting from the highest bin and proceeding backward, according to a procedure based on the following relationships:

$$\mathbf{W}_{smooth}(k) = \sum_{l=1}^{L} \overline{\mathbf{W}}(k+l) h_l(k) \quad (13)$$

$$\mathbf{h}(k) = \mathbf{h}(k+1) + \mu \frac{e(k)^* \mathbf{D}(k)}{\mathbf{D}(k)^H \mathbf{D}(k) + \varepsilon} \quad (14)$$

$$e(k) = |\overline{\mathbf{W}(k)}| - \mathbf{h}(k+1)^H \mathbf{D}(k) \quad (15)$$

where $\mathbf{h}(k)$ is the vector $[h_1(k), h_2(k), ..., h_L(k)]^T$ of the complex-valued coefficients of the smoothing filter evaluated at frequency k, L is the order of the filter, μ is the step-size, ε is the normalization factor and $\mathbf{D}(k)$ is the vector of the observed determinant values:

$$\mathbf{D}(k) = [d_1(k), d_2(k), ..., d_L(k)]^T, \quad d_l = |\overline{\mathbf{W}}(k+l)|. \quad (16)$$

The main steps of the proposed approach are summarized in the following pseudo code:

W$_{smooth} = I$
for *k=highest_frequency_index* ***to*** *1*
    W$_0(k)$=***W***$_{smooth}$
    $L = \min(L_{max}, highest\_frequency\_index - k + 1)$
    compute ***W***(k) *by ICA, starting from* ***W***$_0(k)$
    normalize ***W***(k) *as in (11)-(12)*
    estimate ***W***$_{smooth}$ *as in (13)-(16)*
end

where L_{max} is the maximum order adopted for the smoothing filter. By using such a recursive initialization, ICA is constrained to converge to the solution which guarantees a frequency coherence of the observed separation matrices $\mathbf{W}(k)$. In other terms, the recursion increases the probability that the ratios $r_i(k)$ represent the propagation models of the acoustic waves related to the direct path of the sources. Across the frequencies, the phase coherence of the observed propagation models is increased and consequently the SCT analysis can be performed accurately.

5. EXPERIMENTAL RESULTS

An algorithm of multiple TDOA estimation has been implemented both in Matlab and in C++ and works in real-time on a normal laptop. As a first step a short-time Fourier analysis was performed in order to obtain a frequency-time representation of the observed mixtures $\mathbf{y}(k, \tau)$. For each frequency bin the demixing matrices $\mathbf{W}(k)$ were obtained by applying a Scaled Natural Gradient [7]. The states were obtained as in (5) and (8). The cumulative SCT analysis integrates the euclidean distance between the model and the observed states, across all the time-frequency observations. To perform an online TDOA estimation, rather than integrating all the observed states, the SCT envelopes were recursively averaged in each time-block. The resulting cumulative SCT envelope has been estimated as follows:

$$\overline{cSCT}_b(\tau) = \frac{1}{b} SCT_b(\tau) + \frac{(b-1)}{b} \overline{cSCT}_{b-1}(\tau) \quad (17)$$

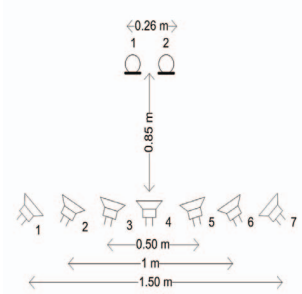
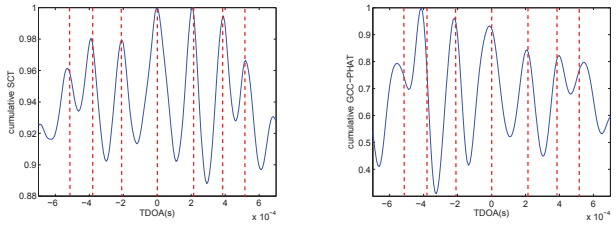


Fig. 1. Experimental setup for the case of 7 sources.



(a) Cumulative SCT profile computed with blocks of 300 ms.

(b) Cumulative GCC-PHAT computed with frames of 4096 samples.

Fig. 2. cSCT and cumulative GCC-PHAT, SNR=20dB (the red dotted lines are the true expected TDOAs).

where b is the time-block for which the $SCT_b(\tau)$ is evaluated. The algorithm is able to detect the TDOAs of many sources. In this experiment we evaluated the case of utterances produced by loudspeakers located as shown in figure 1. In this experiment the algorithm has been evaluated for the estimation of the TDOAs of 7 loudspeakers playing simultaneously sound files of about 10 seconds: 3 male utterances, 3 female utterances, 1 pop song. All the sources were overlapping in time. The resulting average angular distance between the loudspeakers was about 13°. Recordings were performed in a room with $T_{60} = 700\ ms$ with a sampling rate of $fs = 16kHz$ and the FFT analysis was performed with an Hanning window of 2048 samples and a frame-shifting of 512 samples. The length of the time-block used for the ICA and the SCT analysis was $300\ ms$. Since the signals were recorded with two microphones on a distance of $0.26\ m$ of each other, according to the sound speed (e.g. $340\ m/s$) the maximum admissible time-delay was expected to be of about $\pm 0.77\ ms$. Then, the SCT was computed for 180 uniformly spaced values of τ in the range from $-0.77\ ms$ to $+0.77\ ms$.

In our earlier work [2] we showed that under specific condi-

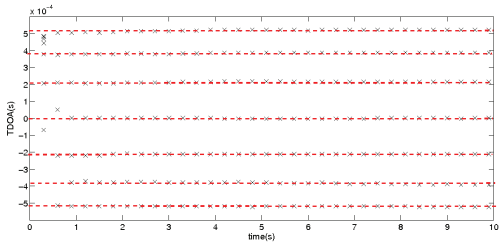
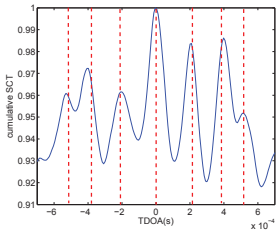
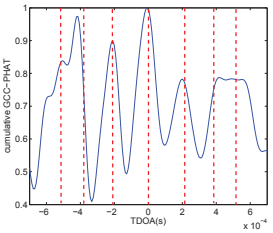


Fig. 3. Selected TDOAs from the estimated cumulative SCT (the red dotted lines are the true expected TDOAs).

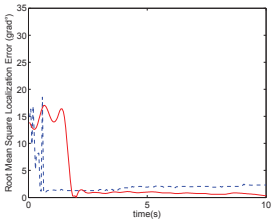


(a) Cumulative SCT profile computed on blocks of 300 ms.

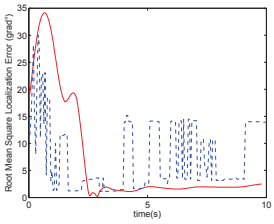


(b) Cumulative GCC-PHAT computed with frames of 4096 samples.

Fig. 4. cSCT and cumulative GCC-PHAT, SNR=5dB (the red dotted lines are the true expected TDOAs).



(a) case of SNR=20dB.



(b) case of SNR=5dB.

Fig. 5. Average Root Mean Square localization error for the DOAs computed with the cumulative GCC-PHAT (dotted line) and with the cSCT method (solid line).

tions the GCC-PHAT [8] is equivalent to the SCT. Hence, the performance of the cumulative SCT was compared to a cumulative GCC-PHAT derived recursively as in (17) where b represents the time-frame where the GCC-PHAT is evaluated. In order to obtain a theoretical resolution of 180 possible time-delays, an interpolation was applied and the GCC-PHAT has been computed and accumulated over frames of 4096 points with step of 256 points. Figure 2 shows the final envelopes associated with the cumulative SCT and with the cumulative GCC-PHAT when the signals are affected by an Additive White Gaussian Noise (AWGN) resulting in a SNR of $20dB$. Both the envelopes shows clear peaks at values close to the corresponding expected TDOAs (red dotted lines). Figure 4 shows the envelopes obtained in case of a $SNR = 5dB$. Note that, even in presence of lower SNR, the cumulative SCT still maintains clear peaks located at the corresponding theoretical TDOAs values. The advantage of the SCT stems from the more accurate estimation of the propagation model since the ICA stage is less sensitive to the noise than the GCC-PHAT. In figure 3 the estimated TDOAs selected by means of the peaks of the cumulative SCT are plotted. One can note that after a few seconds the estimated TDOAs approach the correct values.

Figure 5 compares the Root Mean Square localization Error (RMSE) averaged over all the sources. The corresponding directions of arrival were computed according to the geometrical information and using the TDOAs estimated at each time block. One can observe that for a $SNR = 20dB$ both the cumulative GCC-PHAT and the cumulative SCT converge in few seconds to a small error. Note that the SCT converges to an error very close to the theoretical value ($0.5°$) expected in the case of a resolution of $1°$. Finally, for the case of a $SNR = 5dB$ we observe that the cSCT clearly outperforms the cumulative GCC-PHAT which does not converge to an acceptable error even after 10s of data.

6. CONCLUSIONS

This work introduced a new method to accomplish multiple TDOA estimation by using only two microphones. The cumulative State Coherence Transform and the recursive ICA analysis represent the most relevant components that contributed to derive a robust and effective solution to the problem of multiple speaker localization. Experimental results show that with this approach it is possible to accurately estimate the TDOA of 7 sources, located in a highly reverberant environment, by just analyzing few seconds of data, even in presence of strong noise. The advantages with respect to the use of a cumulative GCC-PHAT were also highlighted. Next activities will concern an investigation on the impact of the proposed solution using tasks commonly adopted for speaker localization benchmarking.

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] F. Nesta, M. Omologo, and P. Svaizer, "A novel robust solution to the permutation problem based on a joint multiple TDOA estimation," in *Proceedings of IWAENC*, Seattle, USA, Sept. 2008.

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

[4] F. Nesta, M. Omologo, and P. Svaizer, "Multiple TDOA estimation by using a state coherence transform for solving the permutation problem in frequency-domain BSS," in *Proceedings of MLSP*, Cancun, Mexico, Oct. 2008.

[5] F. Nesta, P. Svaizer, and M. Omologo, "A BSS method for short utterances by a recursive solution to the permutation problem," in *Proceedings of SAM*, Darmstadt, Germany, July 2008.

[6] Y.S. Choi, H.C. Shin, and W.J. Song, "Robust regularization for normalized LMS algorithms," *IEEE Transaction on circuits and system*, vol. 53, no. 8, pp. 627–631, Aug. 2006.

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

[8] 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.