DOA ESTIMATION OF SPEECH SIGNALS USING SEMI-BLIND SOURCE SEPARATION TECHNIQUES

Ilyas Potamitis, Panos Zervas, Nikos Fakotakis

Wire Communications Laboratory, Electrical and Computer Engineering Dept., University of Patras, 261 10 Rion, Patras, Greece, Tel: +30 2610 991722, Fax: +30 2610 991855
potamitis@wcl.ee.upatras.gr

ABSTRACT

In this paper we investigate the application of complex independent component analysis (ICA) to the direction of arrival (DOA) estimation problem of wideband signals. The ICA based technique is semi-blind in the sense that the structure of the array is known to be uniform and linear (ULA). We show that when the array is ULA the mixing matrix is forced to have the structure imposed by the directivity vectors of the microphone array. ICA is applied to the spectral bins having the higher SNRs. The DOAs are derived from the histogram and clustering of the angle of arrivals of all high SNR spectral bins. The effectiveness of the approach is evaluated on speech signals and is compared against a variety of wideband DOA estimation techniques based on second order statistics. Finally, a toolbox for DOA estimation of wideband sources is made available at: http://slt.wcl.ee.upatras.gr/potamitis/DOA_ICA.zip

1. INTRODUCTION

Angle of arrival estimation has long been of great interest primarily because of the importance of a diverse field of applications it is connected with: DOA estimation is a key component in sonar application, mobile communications (smart antenna array processing), radar applications, microphone arrays as a means for interference reduction and voice separation [1]. The majority of the DOA estimation methods require cross-covariance information among sensors (i.e. ESPRIT, MUSIC, minimum variance), therefore, they are essentially based on second-order statistics [2-3]. Recently, some extensions of these methods were reported based on High Order Statistics (fourth order statistics of the quadricovariance tensor) [4]. The application of ICA to DOA estimation is relatively new and few applications are reported, most of them for narrowband signals [4-7]. ICA is based on information-theoretic criteria to find the mixing matrix (and therefore an estimation of the unmixing matrix) of the statistically independent sources impinging on the microphones [8]. The prime concern of this work is to investigate if the capability of ICA to decorrelate higher moments of the impinging waveforms is beneficial to the process of deriving the DOAs.

The key idea of [6-7] that we extend and validate in this work for wideband signals is that if ICA is applied to the problem of wideband far-field signals impinging on a uniform linear array (ULA), the mixing matrix is forced to have a special structure for each frequency composing the wideband signal. This special structure is imposed by the steering vectors of the sources. Since, in this case, the general form of the mixing matrix is known, there is no amplitude or sign ambiguity as regards the separated sources. This fact is exploited in order to calculate the angles of arrivals of the sources per spectral bin. By following the incoherent approach we calculate the DOAs corresponding to the spectral coefficients having the higher SNR on per block basis [3] and we derive the histogram of all angles of arrival. We validate the ICA-based DOA estimation techniques for a special case of wideband signals, i.e., the case of the speech signal. We compare its performance to second-order techniques in special cases, as when there are limited data available to estimate the angle, at low SNRs, and in the case where the sources have small angular inter-spacing.

2. PROBLEM FORMULATION

2.1 Stating the problem from the beamform perspective

In order to apply beamforming in the context of wideband signals one has to adapt the narrowband techniques to be applicable to wideband, possibly non-stationary signals. To overcome the nonstationary nature of these signals the microphone outputs are segmented into fixed blocks and signal stationarity is assumed for each block and each FFT bin is considered a narrowband signal. The output from a linear microphone array of $M$ sensors receiving $N$ far-field wideband signals sampled at time instants $t=1,...,T$ and bin $f$ is $x(f,t)=[X_1(f,t),...,X_M(f,t)]^T \in \mathbb{C}^{M \times 1}$ where, $X_m(f,t)$ $m=1,...,M$ is the short-time FFT coefficient of the signal received at the $m^{th}$ microphone. The array observation vector is assumed to satisfy:

$$x(f,t)=A(f,\theta)s(f,t)+n(f,t)$$

where,

$$A(f,\theta)=[a_1(f,\theta),...,a_M(f,\theta)]^T \in \mathbb{C}^{M \times N}$$

$$s(f,t)=[s_1(f,t),...,s_N(f,t)]^T \in \mathbb{C}^{N \times 1}$$

$$n(f,t)=[n_1(f,t),...,n_M(f,t)]^T \in \mathbb{C}^{M \times 1}$$

the matrix $A(f,\theta)$ is referred to as the steering matrix having as columns the steering vector of each source $n$ in frequency $f$ from direction $\theta_n$ and $a_m(f,\theta_n)=[1,e^{jkd\cos(\theta_n)},...,e^{jkdM\cos(\theta_n)}]^T$. Let $L$ the window size and also the size of the FFT then $f,t,L,l=0,...,L-1$ where $f$ is the sampling frequency, $k=2\pi f/c$, $d$ microphone spacing, $c$ sound velocity ($c=343$ m/sec).
2.2 Stating the problem from the ICA perspective

Blind signal separation (BSS), has Independent Component Analysis (ICA) as its main representative [8]. In the context of speech separation, ICA is based on super-Gaussian probabilistic models of the sources and information-theoretic criteria in order to find a linear transform which, when applied to the mixture of the voices, returns statistically independent sources. ICA assumes that each microphone receives a linear mixture of the sources. We retain the same notation as in par. 2.1

\[ x(f,t) = B_f s(f,t) + n(f,t) \]  
(5)

If one applies complex independent component analysis on \( x(f,t) \) for every \( f \) and the array holds an ULA structure, then the mixing matrix discovered must be equal to \( A(f) \) of (1). Since the separating matrix discovered is a permuted and scaled version of the mixing matrix, then

\[ A(f) = B P D \]

where \( D \) is a full-rank diagonal matrix, which, when applied to the separating matrix \( B \), must return the \( A(f) \) form of (2). \( P \) is a permutation matrix acting on the columns of \( B \). One can easily recognize that \( D \) is bound to have the form:

\[ D = \text{diag}(1/b_1, ..., 1/b_N) \]

where \( b_i \) is the first entry of source \( i, i=1,...,N \). The permutation ambiguity refers only to column permutation corresponding to which source is estimated first, and for our application does not need to be resolved. Permutation matrices can be different for different frequency bins but the task is DOA estimation and not BSS. The DOA estimation procedure based on peak-picking the histogram of DOAs derived from the selected frequency bins (see section 4) is not affected by the different permutation matrices among spectral bins.

One can observe that the ULA imposes a constraint on the estimation procedure of complex ICA which has several implications:

a) The application of ICA to speech separation of voices in real situations is complicated due to the movement of the speaker. The movement implies a time-varying mixing matrix which, in turn, restrains the number of available signal samples from which the demixing coefficients are estimated. In the proposed formulation the separation problem of \( n \) sources with \( n \) microphones the mixing matrix to be estimated by any ICA algorithm is \( n^2 \), whereas, in our case, in theory, it can be reduced to just \( n \) (corresponding to the number of the sources).

b) The sign and scale ambiguity problem does not arise due to the special form of the mixing matrix.

In order to further increase the robustness of the estimation procedure while decreasing the computational load we enforce the calculation of DOAs to be derived only from the frequencies possessing the highest SNR on per block basis. Spectral bins with low SNRs are mainly due to noise or they are frequently corrupted by noise and thus tend to return fairly non-robust DOA estimates.

3. COMPLEX DOA ESTIMATION

The DOA estimation procedure, in principle, does not require a specific ICA technique as long the latter can be applied to complex time-series. Therefore, a variety of algorithms can be directly applied to the problem as [9-11], (to mention a few well known ICA techniques that have variations that can be applied to complex time-series). The best results however were achieved by using JADE (Joint Approximate Diagonalization of Eigen-matrices) [11] a fact also reported in [6]. The key points of JADE which is based on 4th order cumulants is the joint diagonalization of a set of matrices (in our case complex). The joint diagonalization of a set of square matrices consists in finding the orthonormal change of basis which makes the matrices as diagonal as possible. We do not go in much detail because the algorithm is well documented [11]. A short description of the technique is as follows:

1) Estimate a whitening matrix \( W \) and set

\[ z(f,t) = W x(f,t) \]

2) Form the \( 4^{th} \)-order cumulant tensor and compute its \( n \) most significant eigenmatrix pairs \( (\lambda_i, M_i) \).

3) Jointly diagonalize the \( n \) matrices \( \lambda_i M_i \) to calculate the orthonormal matrix \( V_f \) (the rotation matrix), such that the cumulant matrices are as diagonal as possible. The matrix \( V_f \) is found by optimizing an orthogonal contrast

\[ \min_{V_f} \sum_{i \neq j} |V_f M_i V_f^H| \]

where \( off \) of a matrix \( N \) is defined as: \( off(N) = \sum_{i \neq j} |n_{ij}|^2 \)

4) The estimate of \( A(f,t) = W V_f \).

The matrix \( A(f,t) \) is derived for every spectral line.

4. DESCRIPTION OF THE ALGORITHM

The majority of the DOA estimators is derived under the assumption of the narrowband model and subsequently extended to wideband emitters such as the speech signal. In our case, we follow the incoherent approach; therefore the wideband array output of each microphone is decomposed into narrowband components by using discrete Fourier transform (DFT). Subsequently, the complex ICA transform is applied to the narrowband components and the narrowband DOA estimates are averaged. Let \( M \) be the number of microphones, \( N \) the number of wideband sources and \( L \) the FFT size.

**Step 1**: Select the \( k \) frequency bands of the DFT transformed current block corresponding to the narrowband components with the largest power values.

**Step 2**: For each segment of \( T \) frames, construct a matrix \( X \in \mathbb{C}^{M \times T} \), where each row of \( X \) holds the complex time-series of each microphone at the currently selected strongest frequency of the \( k \) set.

**Step 3**: Apply Jade to \( X \) as described in PAR. 3 and derive the mixing matrix \( B_k \in \mathbb{C}^{M \times N} \), Divide each column with the first element of the column to resolve sign/amplitude ambiguity.
Step 4: Take the angle of the complex numbers composing $B_k$. Let $b_i$ be the second row of $B_i$ and $i$ denote the columns index corresponding to the sources.

Step 5: Calculate DOA from:

$$\theta = \arccos \left( \frac{c}{2 \pi f_b} \right) \cdot \frac{b - 1}{L}, \quad f_b = \frac{f_{\text{LOF}}}{L}, \quad b = 1 \ldots L/2 + 1, \quad i = 1 \ldots N,$$

where the parameters of the implementation are shown in Table I.

Steps 2-4 are repeated for each selected frequency $k$. Average the DOA estimates over all $k$ frequency bands and take the histogram of the data. The angles corresponding to the $N$ largest peaks of the histogram are the estimated DOAs of the speech sources.

A flow diagram of this process is depicted in Fig. 1.

5. SIMULATIONS AND RESULTS

In order to validate the proposed technique for wideband signals we choose speech signals. However, since we do not make use of any special characteristics of the speech signal the subsequent analysis holds for any kind of wideband signal.

In general, the quality of the DOA measurements corresponding to speakers’ angles of arrival is affected by the following three factors:

a) The spectral content of the speech segment used to derive the DOA. An utterance is composed of a succession of high energy – low energy segments interspersed with silence parts. The low energy speech and silence parts are sensitive to background noise and are likely to return erroneous DOA measurements.

b) The reverberation level of the room can give rise to spurious measurements due to reflections on the walls and objects of the enclosure.

c) The relative positioning of the array with respect to the talkers, the number of simultaneous sources present in the receptive field and their relative positioning.

One needs to design a detailed experimental procedure in order to rank comparative performance results between ICA-based wideband DOA estimation and second order techniques. In our simulations, due to limited space, we focus on a) one speaker case under Gaussian corruption, b) two speakers’ case, and c) one speaker case for various block sizes of available data to derive the DOAs.

In the one speaker case the speech signal is corrupted by additive white Gaussian noise (AWGN) interference at various SNRs (Table II). In our experiments, the results are averaged over 100 Monte Carlo simulations for each SNR ranging from 0 dB to clean speech. For each simulation a random true angle between 1 and 179 degrees is chosen. The broadband signals used are speech recordings chosen randomly from the TIMIT database and concatenated to form 30 seconds of speech. From Table II and extensive experimentation we observed no bias in the estimation of the DOA, while a graceful increase of the standard deviation is observed for decreasing SNR. For 0 SNR and below the DOA estimation procedure becomes unstable and there can be estimated completely erroneous angles that affect both bias and variance of the angles measurements. This effect is prominent for angles close to the end-fire position.

In Fig. 2, we depict comparative results between the proposed method and minimum variance and delay and sum for a specific case where there is evident sparsity of data to derive the available information. For a single speaker, the ICA based technique compares almost equally to all second-order techniques we employed, including wideband MUSIC and wideband ESPRIT. However, we did not observe a distinct advantage of the proposed method against second-order methods which proved more robust under AWGN interference. The ICA-based technique had some advantage when the DOA was derived from a limited number of overlapping windows.

Furthermore, for more than one speaker, the ICA-based wideband DOA estimation technique did not compare favourably to second-order techniques especially compared to MUSIC and ESPRIT, except for some very specific situations. Numerous experiments show that in specific situations, as e.g., when speakers have small angular spacing and the number of samples to derive the DOAs is small (as it would be in the case of a moving speaker), the ICA-based technique can achieve better results (see also Fig. 3).

![Diagram](image)

**Table I**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFT and window size</td>
<td>N=512 samples</td>
</tr>
<tr>
<td>Overlap</td>
<td>256 samples</td>
</tr>
<tr>
<td>Sampling freq.</td>
<td>$f_s=8000$ Hz</td>
</tr>
<tr>
<td>No. of high SNR bands</td>
<td>128</td>
</tr>
<tr>
<td>Window</td>
<td>Hamming</td>
</tr>
<tr>
<td>Speed of sound</td>
<td>$c=343$ m/sec</td>
</tr>
<tr>
<td>Microphone separation</td>
<td>$d=0.01$ m</td>
</tr>
<tr>
<td>Pre-emphasis applied</td>
<td>$H(z)=1 -0.95 z^{-1}$</td>
</tr>
</tbody>
</table>

Parameters of the ICA based implementation.
Table II

<table>
<thead>
<tr>
<th>SNR</th>
<th>Mean (degrees)</th>
<th>Std. (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>0.15</td>
<td>0.01</td>
</tr>
<tr>
<td>20</td>
<td>-0.07</td>
<td>0.15</td>
</tr>
<tr>
<td>15</td>
<td>-0.05</td>
<td>0.77</td>
</tr>
<tr>
<td>10</td>
<td>-0.65</td>
<td>3.19</td>
</tr>
<tr>
<td>5</td>
<td>-0.61</td>
<td>7.60</td>
</tr>
<tr>
<td>0</td>
<td>3.45</td>
<td>26.54</td>
</tr>
</tbody>
</table>

Single speaker, fixed position. Wideband DOA estimation by ICA based technique. 100 Monte-Carlo runs.

Fig. 2 The true DOA is at 50°. DOA estimation by Delay & Sum, Minimum Variance and ICA based techniques. All algorithms used only 5 overlapping frames.

Fig. 3 The true DOAs are at 20° and 25°. DOA estimation by Delay & Sum, Minimum Variance and ICA based techniques. All algorithms used 10 overlapping frames.

Extensive comparative experimentation did not establish a consistent advantage of the ICA-based technique for DOA estimation of wideband signals compared to second-order techniques. We are currently investigating the reasons of this failure as the experimental results are not conclusive. In the case of few available frames this failure may be due to the inability to achieve independence from sparse data and allow for an accurate estimation of the mixing matrix from which the angles of arrival are derived.

6. CONCLUSIONS

We presented a method for estimating the angles of arrival of wideband sources using complex independent component analysis and a uniform linear array structure. We have shown that in the case of a single wideband source, the ICA-based DOA estimation compares equally and in certain situations surpasses popular second-order techniques. However, we did not have enough evidence to establish the superiority of the ICA based technique in all situations. We are currently investigating the reasons of this as well as its performance in reverberating environments. Although the proposed technique was validated by using speech signals we did not make use of any information concerning the spectral structure of speech, and, therefore, this approach is expected to be directly applicable to the spectral signatures of moving vehicles and low-flying aircrafts tracked by uniform, linear microphone arrays.

ACKNOWLEDGEMENT

We acknowledge the financial support provided by the EC project INSPIRE: IST-2001-32746.

REFERENCES