MULTI-SOURCE LOCALIZATION IN REVERBERANT ENVIRONMENTS

Elio D. Di Claudio(*) , Raffaele Parisi(**) and Gianni Orlandi(***)
INFOCOM Dpt., University of Rome “La Sapienza”
Via Eudossiana 18, I-00184 ROMA RM Italy
(*) Tel: +39-06-44585837; fax +39-06-44585490
e-mail: dic@infocom.uniroma1.it
(**) Tel: +39-06-44585486; fax +39-06-44585490
e-mail: parisi@infocom.uniroma1.it
(***) Tel: +39-06-44585490; fax +39-06-44585490
e-mail: orl@infocom.uniroma1.it

ABSTRACT
The very large relative bandwidth of acoustic sources, coupled with the high number of reflections of a typical listening room, makes localization a challenging task, since all basic assumptions of classical array processing algorithms constitute at the best viable approximations in real-world environments. In this work, a novel decentralized approach for acoustic localization in reverberant environment is presented. It is based on a two-stage strategy. First, candidate source positions are found by a Time-Delay-Of-Arrivals (TDOA) analysis of signals received by colocated pairs of microphones. Differential delays are estimated by a robust ROOT-MUSIC based technique, applied to the sample cross-spectrum of whitened signals recorded from each microphone pair. A subsequent clustering stage in the spatial coordinates validates the raw TDOA estimates, eliminating most of false detections. The new algorithm is capable of tracking multiple speakers at the same time, exhibits a very good consistency of location estimates, and compares favourably with previous approaches.

1 INTRODUCTION
Localization of acoustic sources in reverberant environments is an important task in many automatic systems for surveillance, videoconferencing, hands-free talking [1]. Spatial parameters obtained in the localization process can be used in a variety of applications: dereverberation of speech, fault prediction and analysis in machinery, cueing and tracking of TV cameras, speaker verification, etc.

From a signal processing standpoint, the issue is how to deal with multiple arrivals and distinguish among useful signal(s) and reverberations. The presence of reflective surfaces in closed environments is usually modeled by virtual sources [2], whose number typically exceeds the microphone array size. This fact, coupled with the very large bandwidth of signals of interest, makes unsuitable the parametric techniques developed for narrow-band or moderately wide-band array processing in the presence of far-field sources [3][4][5]. This is the reason why most approaches involve the estimate of differential time delays (Time Delay of Arrival, TDOA) among pairs (doublets) of co-located microphones, to get an efficient and reasonably cheap localization [6][7][8][9]. Most algorithms for TDOA estimation in the presence of strong multipath are based on the generalized cross-correlation function [6][9]: usually, a single-source model is assumed and a joint parameter optimization from signals collected by many sensors at a time is performed.

From the point of view of system design, it is very important to reduce synchronization requirements and signal paths to a minimum, in order to allow a decentralized and asynchronous processing of TDOAs.

In this work, we propose a novel two-stage strategy for the robust localization of multiple speakers in reverberant rooms. The first stage computes the TDOAs for the direct path and early (strongest) reflections by a closed-form parametric approach, based on the ROOT-MUSIC algorithm [10] and a proper data preprocessing.

The second stage finds the most likely position of the speakers by means of a clustering in space performed among all the estimated locations. The most dense clusters are selected as candidate speakers, thus eliminating most of false detections generated by outliers (virtual sources, localization ambiguities, impulsive noise, etc.).

It is shown that location estimates obtained are characterized by a very low bias and variance, if compared to existing approaches, which enables a very simple and reliable clustering.

2 PROPOSED ALGORITHM
In the following, we briefly describe the main steps of the proposed approach.

2.1 Signal whitening through LPC
Array microphones are paired in doublets. The two microphones of each doublet must be sufficiently close to
each other and their electrical responses should be matched enough, so that they essentially receive a time-shifted copy of each source signal.

Since the range of the speech spectrum is quite high in acoustic applications, it can heavily influence the quality and robustness of the subsequent TDOA estimation [6][7][9]. For this reason, signals coming from each doublet are pre-processed by a standard LPC algorithm, that removes the common spectral features, including the pitch, and minimizes the spectrum range. The short- and long-term predictors [11] are computed on the basis of the average autocorrelation of doublet signals. Pre-whitening is an essential feature of optimal (and robust) TDOA estimation [6]. The proposed preprocessing avoids the requirement for compensating the differential group delay between pre-whitening filters, which occurs when they are computed separately for each microphone. The average LPC prediction is also more stable from a statistical point of view, since twice the number of samples are involved for the estimate of the same parameters, resulting in smoother transfer functions. It should also be realized that the spectrum of both signals from each doublet should be nearly the same for the assumptions made above. In the presence of multiple speakers, more long-term prediction steps can be applied sequentially, until residuals have been sufficiently whitened.

2.2 Robust parametric TDOA algorithm

In alternative to approaches based on the generalized cross-correlation, the TDOA search can be recasted as a disturbed harmonics retrieval problem [12] in the frequency domain. In fact, indicating with $x_i(t)$ and $x_j(t)$ the signals acquired by the generic doublet, it can be shown that the following equation holds for the cross-power spectrum $P_{1,2}(f)$, under mild hypotheses:

$$P_{1,2}(f) = E\left[ X_1(f)X_2(f) \right] = \sum_{p=1}^S S_p(f) \sum_{k=1}^{K_p} \alpha_{pk}(f) \alpha_{p'k'}(f) e^{-j2\pi(f_\tau_k - f_\tau_{k'})} \tag{1.1}$$

where $S$ is the number of speakers, $K_p$ is the number of significant arrivals from a single speaker (direct path and reflections), $S_p(f)$ is the average cross-spectrum of the whitened speech at the doublet centroid, $\alpha(f)$ are slowly-varying transfer functions, and $(f_\tau_k - f_\tau_{k'}) = \Delta_{\tau kl}$ are the TDOAs to be estimated. $P_{1,2}(f)$ can be interpreted as a sum of few sinusoids, modulated by slowly-varying envelopes and embedded in a nearly white noise [13]. This assumption is supported also by empirical evidence. Larger envelope fluctuations are present near formant frequencies (zeros of the prediction filter) and can be regarded as noise.

The ROOT-MUSIC algorithm [10], which is very robust to envelope fluctuations, has been successfully used to estimate the sinusoid frequencies. $P_{1,2}(f)$ has been estimated from consecutive frames of LPC residuals by means of unwindowed FFT and time averages.

Toeplitz (or Hankel) data matrices are built with the discrete samples of the cross-spectrum. Pre-whitening largely improves the numerical balance and conditioning of resulting data matrices.

The effective rank of the data matrix, which is representative of the number of significant arrivals, has been estimated by a new robust algorithm [16]. It is assumed that the smallest eigenvalues of the covariance matrix are clustered around the noise power with a distribution which is approximately Gaussian or $\chi^2$-Square. In contrast, “signal” eigenvalues [10] belong to a different and unknown distribution [17] and can be regarded as outliers when observed on the eigenvalue spectrum under a gross error modeling [16]. This innovative interpretation allows to use robust estimators of the noise eigenvalue distribution to set the threshold among signal and noise eigenvalues [16]. In particular, the sample median and the absolute median deviation of eigenvalues have been successfully used in this work.

For median-based estimators, it is known that the order of the data matrix must exceed twice the number of significant reflections [16]. Information theoretic criteria like MDL or AIC [17] do not have this limitation but they have been discarded for the excessive sensitivity to modelling errors and approximations.

2.3 Source clustering

From two TDOAs estimated from each pair of doublets, a candidate source position or a bearing line is extracted by efficient geometric algorithms [14]. Estimates falling out of room borders are immediately discarded. The remaining estimates are clustered in space, using a fast fuzzy algorithm [15].

The maximum number $L_p$ of estimates that can be attributed to a single speaker is determined by the array geometry. For example, in the case of 2D-estimation:

$$L_p = \frac{n}{2} K \tag{1.2}$$

where $n$ is the number of doublets and $K$ is the number of TDOA sets used for the clustering. False detections due to geometrical ambiguities, wrong pairing of TDOA estimates and reverberations generate disperse clusters collecting few locations. In contrast, dense clusters having an approximately elliptical shape are obtained around speakers. Finally, centroids of those clusters gathering a number of elements close to $L_p$ are selected as speaker locations.

3 EXPERIMENTAL RESULTS

The ROOT-MUSIC TDOA algorithm was compared with the Cross-Power Spectrum Phase approach (CPSP, [9]) using simulated data generated by the image method.
Results herein described refer to a room of size (6x7x4 m); the reflection coefficients were $\beta = 0.8$ for the walls and $\beta = 0.6$ for the ceiling and the floor. Four square arrays of size 4 placed on the walls were considered. In particular, air absorption was simulated by selecting an attenuation coefficient of 1 dB/m at 20 KHz.

Pre-whitening was found to improve the performance of both algorithms. Table 1 shows the sample TDOA statistics for a single speaker, obtained from one doublet. Figure 1 demonstrates the tracking abilities of ROOT-MUSIC in the presence of three simultaneous speakers. In abscissas the temporal frame index is reported; 1024 samples are collected for each frame at 44,1 KHz sampling frequency. TDOAs are estimated every frame. Tracks correspond to the polynomial roots, ranked with respect to the smallest distance from the unit circle [10]. Figures 2 and 3 show an example of clustering over time in the presence of three simultaneous speakers. Spatial coordinates are normalized to $cT$, where $c$ is the acoustic wave propagation speed and $T$ is the sampling period.

References


<table>
<thead>
<tr>
<th>TDOA Estimator</th>
<th>sample bias and standard deviation (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNR = 10 dB</td>
</tr>
<tr>
<td>CPSP</td>
<td>0.0278, 0.165</td>
</tr>
<tr>
<td>Root-Music</td>
<td>0.0126, 0.0472</td>
</tr>
</tbody>
</table>

Table 1: TDOA estimator performance comparison for prewhitened signals.
Fig. 1: TDOA tracking of three speakers by ROOT-MUSIC TDOA.

Fig. 2: Clustering example; x-marks indicate raw location estimates, circles point to true speaker positions.

Fig. 3: Scatter diagram of final location estimates over time, circles point to true speaker positions.