MULTI-SOURCE LOCALIZATION IN REVERBERANT ENVIRONMENTS BY ROOT-MUSIC AND CLUSTERING

E. D. Di Claudio, R. Parisi and G. Orlandi

INFOCOM Dpt.
University of Rome "La Sapienza"
Via Eudossiana 18, I-00184 ROMA RM Italy

ABSTRACT

Localization of acoustic sources in reverberant environments by microphone arrays remains a challenging task in audio signal processing. As a matter of fact, most assumptions of commonly adopted models are not met in real applications. Moreover, in practical systems it is not convenient or possible to employ sophisticated and costly architectures, that require precise synchronization and fast data shuffling among sensors.

In this paper, a new robust multi-step procedure for speaker localization in reverberant rooms is introduced and described. The new approach is based on a *disturbed harmonics* model of time delays in the frequency domain and employs the well-known ROOT-MUSIC algorithm, after a preliminary distributed processing of the received signals. Candidate source positions are then estimated by clustering of raw TDOA estimates.

Main features of the proposed approach, compared to previous solutions, are the capability of tracking multiple speakers and the high accuracy of the closed form TDOA estimator.

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 a proper treatment of multiple arrivals, corresponding to both useful signal(s) and reflections. Reflective surfaces in closed environments are usually modeled by the introduction *of virtual sources* [2], whose number typically exceeds the microphone array size. This fact, coupled with the very large bandwidth of the signals of interest, makes unsuitable the parametric techniques used in narrow-band or moderately wide-band array processing in the presence of far-field sources [3][4][5]. For these reasons, most approaches to source localization involve the use of differential time delays (Time Delay of Arrival, TDOA) among pairs ("doublets") of co-located microphones [6][7][8][9]. This process requires a joint parameter optimization from signals collected by many sensors at a time.

Typically, TDOA estimation is performed by *generalized cross-correlation* methods [6][9], that are appealing for their simplicity and ease of implementation. Anyway, generalized cross-correlation methods assume a single-source model, which can be far from reality in many typical operating environments. A different model and strategies are thus needed to overcome the limitations of traditional approaches.

From the point of view of system design, it is very important to reduce synchronization requirements and signal paths to a minimum, to reduce costs in current applications.

In this work, we propose a novel three-stage strategy for the robust localization of *multiple* speakers in reverberant rooms. The first stage consists of *data pre-whitening* by use of Linear Predictive Coding (LPC). The effects of signal pre-whitening are to generate an approximate concentration of the likelihood function (under a simplifying Gaussian assumption) [7] and to reduce the reverberation effects (e.g. the number of significant TDOA to be estimated) .

In the second stage the TDOAs for the direct path and early (strongest) reflections are estimated by a closed-form parametric approach, based on the ROOT-MUSIC algorithm [12].

Finally, the third 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.).

2. PROPOSED APPROACH

In this section we briefly describe the main steps of the proposed approach.

2.1 Signal whitening through LPC

Array microphones are paired in doublets. Microphones in each doublet are supposed to be physically close to each other, so that

they can be assumed to receive a time-shifted replica of each source signal, filtered by the same acoustic transfer function.

Typically, the range of the signal spectrum in acoustic applications is characterized by high fluctuations ($10^4 \div 10^6$). This fact can heavily influence the quality and robustness of the subsequent TDOA estimation [6][7][9]. For this reason, in the proposed approach signals coming from each pair of microphones are pre-processed by a standard LPC algorithm, which has been recently applied to speech enhancement, also in reverberant environments [10][11]. The main effect of LPC is to remove the common spectral features present in doublet signals (including the pitch) and minimize spectrum fluctuations, induced by multipath. Whitening produces data matrices that are better balanced, well conditioned and largely insensitive to speech nonstationarity, thus enhancing the subsequent parametric TDOA processing. Being a linear processing, LPC does not modify either the general signal model (described in the following section), or the local signal-to-noise ratio (SNR) in the frequency domain.

Short- and long-term predictors [13] are computed on the basis of the *average autocorrelation* of doublet signals. The filter resulting from the average of the two auto-correlation functions is smoother in frequency and leads to a negligible loss in performance with respect to optimal processing [6]. In contrast, separate filtering of each signal would require compensation of the different group delay during the estimation of the phase spectrum, leading to an increased computational cost.

2.2 Robust parametric TDOA estimation by ROOT-MUSIC

In alternative to approaches based on the generalized cross-correlation, the TDOA search can be also recasted as a *disturbed harmonics* retrieval problem [14][15] from the cross-power spectrum $P_{12}(f)$. If $x_1(t)$ and $x_2(t)$ are the signals acquired by the generic doublet and $X_1(f)$ and $X_2(f)$ are the corresponding Fourier tranforms, the cross-power spectrum $P_{12}(f)$ between $x_1(t)$ and $x_2(t)$ is defined by the following formula:

$$P_{12}(f) = E\left[X_1(f) X_2(f)^* \right] \tag{1.1}$$

where $E[.]$ indicates the expectation operator and $(.)^*$ is the complex conjugation operator.

Under mild hypotheses, it can be shown that the following equation holds for $P_{12}(f)$:

$$P_{12}(f) = \sum_{p=1}^{S} S_p(f) \left(\sum_{k=1}^{K_p} \sum_{l=1}^{K_p} \alpha_{1pk}(f) \alpha_{2pl}^*(f) e^{-j2\pi f(\tau_{1pk} - \tau_{2pl})} \right) \tag{1.2}$$

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 spectrum of the whitened signal at the doublet centroid, $\alpha(f)$ are slowly-varying transfer functions, and $(\tau_{1pk} - \tau_{2pl}) = \Delta_{pkl}$ are the TDOAs to be estimated. Based on this formula, $P_{12}(f)$ can be interpreted as a sum of a few sinusoids, modulated by *slowly-varying* envelopes and embedded in a *nearly white* noise [15]; this assumption is supported also by empirical evidence. $P_{12}(f)$ has been estimated at discrete and equispaced frequencies $\{f_k,\ k = 1,2,\ldots,\ M\}$ from consecutive frames of LPC residuals by means of FFT and time averaging.

The ROOT-MUSIC algorithm [12], which is known to be very robust to envelope fluctuations, has been used to estimate the sinusoid frequencies in eq. (1.2). Namely, ROOT-MUSIC is applied to the covariance of a Hankel-structured data matrix, formed by estimates of $P_{12}(f_k)$ at each doublet [15].

Compared to existing approaches, location estimates obtained with the proposed approach are characterized by a very good consistency. Moreover, since strict synchronization and fast data transfer requirements are confined within each doublet, the proposed strategy enables a decentralized and asynchronous processing of TDOAs by use of DSP.

2.3 Estimation of the number of relevant reflections

The ROOT-MUSIC algorithm requires the estimation of the number of sinusoids from the effective rank of the data covariance matrix, which is representative of the number of significant arrivals and can be used to detect contributions in the received signals due to the direct paths. In the present approach the rank has been estimated by setting a threshold on the eigenvalues of the data covariance via a simple and robust algorithm.

It is assumed that the smallest eigenvalues are clustered around the noise power with a distribution which is approximately Gaussian or χ-Square. In contrast, "signal" eigenvalues [5] belong to a different and unknown distribution [17] and can be regarded as *outliers* when observed on the eigenvalue spectrum.

This 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 the 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 have been discarded for the excessive sensitivity to modelling errors and approximations.

2.4 Source clustering

From TDOAs generated for each pair of doublets, a candidate source position is computed by efficient geometric algorithms available in literature [18]. Estimates that lie inside the room borders are clustered in space, using a fast *fuzzy* algorithm [19].

The maximum number L_p of estimates that can be attributed to a single speaker can be easily computed from the array geometry [18]. Most of remaining estimates are generated by the pairing of TDOAs that refer to different acoustic sources and have no physical meaning. Other incorrect locations arise from geometrical ambiguities of the array, reverberation, diffractions.

While wrong estimates usually generate disperse clusters containing few points, dense clusters having an approximately elliptical shape are formed around the speakers. Centroids of

clusters gathering a number of elements close to L_p are finally selected as speaker locations. Performance can be improved by extending the clustering phase over consecutive time frames.

3. EXPERIMENTAL TRIALS

The ROOT-MUSIC TDOA algorithm was compared with the Cross-Power Spectrum Phase approach (CPSP, [9]) using simulated data generated by the image method [2] and different array configurations. Results herein described refer to a room of size (6x7x4 m), with value β=0.8 for the reflection coefficient of the walls and β=0.6 for the ceiling and the floor. Four square arrays of size 4 place on the walls were considered.

Prewithening was found to improve the performance of both algorithms in any case. Table 1 shows the sample TDOA statistics obtained from a single doublet in the presence of one speaker, for different values of SNR. The better performance of the proposed approach are clearly visible.

Figure 1 demonstrates the tracking capabilities of ROOT-MUSIC in the presence of three simultaneous speakers. Tracks correspond to the polynomial roots, ranked with respect to the smallest distance from the unit circle [12]. Finally, figure 2 shows an example of clustering over time in the presence of three simultaneous speakers.

4. SUMMARY

A new approach to speaker localization in reverberant environments has been introduced and described. The better performance of the proposed method with respect to traditional techniques have been demonstrated in realistic computer simulations, even in the presence of multiple speakers. The computational cost of the new procedure, being based on standard and optimized building blocks (LPC, FFT, small-size SVD, polynomial root finders), is affordable by modern DSP processors and depends on the particular implementation. The final fuzzy clustering can be implemented in high-level languages on general-purpose workstations. Finally, the proposed algorithm is characterized by a high degree of flexibility and can easily incorporate any enhanced whitening, TDOA estimation and clustering techniques.

5. REFERENCES

[1] M. Omologo, P. Svaizer and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Communication*, Vol. 25, No. 1-3, Aug. 1998, pp. 75-95.

[2] J. B. Allen and A. Berkeley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, Vol. 65, No. 4, April 1979, pp. 943-950.

[3] Y. Grenier, "Wideband source location through frequency-dependent modeling," *IEEE Transactions on Signal Processing*, Vol. 42, No. 5, May 1994, pp. 1087-1096.

[4] P. M. Schultheiss, and H. Messer, "Optimal and suboptimal broad-band source location estimation," *IEEE Trans. on Signal Processing*, Vol. 41, No. 9, Sept. 1993.

[5] H. Wang, and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-33, No. 4, Aug. 1985, pp. 823-31.

[6] C. H. Knapp, and G. Clifford Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoust., Speech and Signal Processing*, Vol. 24, No. 4, Aug. 1976.

[7] G. Clifford Carter, *Coherence and time delay estimation*, IEEE Press, 1993.

[8] B. Champagne, S. Bedard and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 2, March 1996.

[9] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No. 3, May 1997.

[10] B. Yegnanarayana, C. Avendano, P. Satyanarayana Murthy, H. Hermansky, "Enhancement of reverberant speech using LP residual," *Proc. of the IEEE Int. Conf. on ASSP* (ICASSP'98), Vol. 1, Seattle, Washington, USA, May 12-15, 1998, pp. 405-408.

[11] B. Yegnanarayana, C. Avendano, H. Hermansky, P. Satyanarayana Murthy, "Speech enhancement using linear prediction residual, " *Speech Communication* 28, 1999, pp. 25-42.

[12] B. D. Rao, K. V. S. Hari, "Performance Analysis of Root-Music". *IEEE Trans. Signal Processing*, Vol. 37, No. 12, December 1989.

[13] R. P. Ramachandran, P. Kabal, "Pitch prediction filters in speech coding," *IEEE Trans. on Acoust., Speech and Signal Processing*, Vol. 37, No.4, Apr. 1989, pp. 467-478.

[14] W.A. Gardner, *Statistical Spectral Analysis, A Nonprobabilistic Theory*, Prentice Hall, 1988.

[15] M. A. Hasan, M. R. Azimi-Sadjadi, and G. J. Dobeck. "Separation of Multiple Time Delays Using New Spectral Estimation Schemes", *IEEE Trans. Signal Processing*, Vol. 46, No. 6, June 1998.

[16] P.J. Huber, *Robust Statistics*, John Wiley, New York, 1981.

[17] M. Wax, T. Kailath, "Detection of Signals by Information Theoretic Criteria," *IEEE Trans. on Acoustic Speech and Signal Processing*, Vol. 33, No. 2, pp 387-392, April 1985.

[18] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A Closed-Form Estimator for use with Room Environment Microphone Arrays," *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No. 1, Jan. 1997, pp. 45-50.

[19] S. Chiu, "Fuzzy Model Identification Based on Cluster Estimation," *Journal of Intelligent & Fuzzy Systems*, Vol. 2, No. 3, 1994.

TDOA Estimator	**sample bias** and **standard deviation** (ms)		
	SNR = 10 dB	SNR = 20 dB	SNR = 30 dB
CPSP	0.0278, 0.165	0.0153, 0.0264	0.015, 0.025
Root-Music	0.0126, 0.0472	0.0087, 0.0183	0.0065, 0.0151

Table 1: TDOA estimator performance comparison (prewhitened signals)

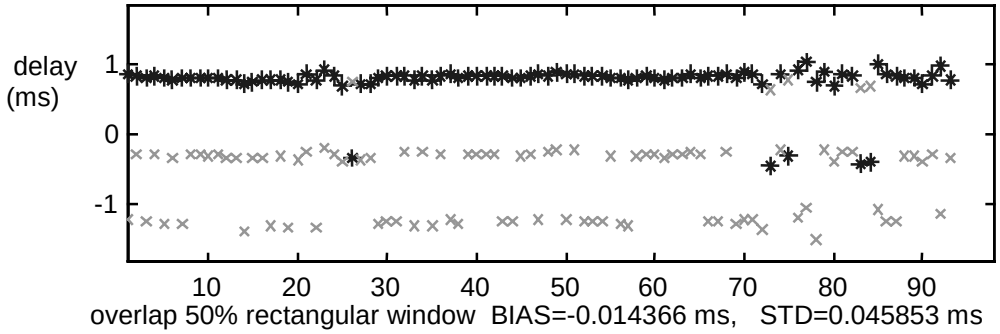

Fig. 1: TDOA tracking of three speakers by ROOT-MUSIC TDOA

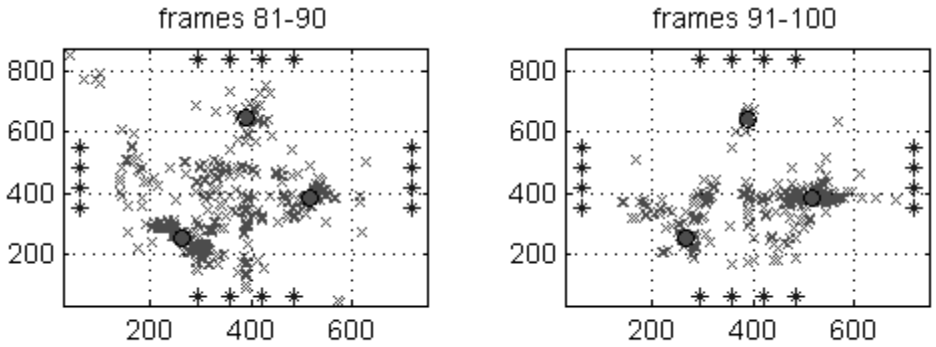

Fig. 2: Clustering example; x-marks indicate raw location estimates, circles point to final speaker estimates (units are in number of cT, where c is the propagation speed of sound and T is the sampling period)