Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Signal Processing 85 (2005) 177–204

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

Time difference of arrival estimation of speech source in a noisy and reverberant environment

Tsvi G. Dvorkind[a,*], Sharon Gannot[b]

[a]Faculty of Electrical Engineering, Technion, Technion City, 32000 Haifa, Israel
[b]School of Electrical Engineering, Bar-Ilan University, 52900 Ramat-Gan, Israel

Received 30 October 2003; received in revised form 23 September 2004

Abstract

Determining the spatial position of a speaker finds a growing interest in video conference scenarios where automated camera steering and tracking are required. Speaker localization can be achieved with a dual-step approach. In the preliminary stage a microphone array is used to extract the *time difference of arrival* (TDOA) of the speech signal. These readings are then used by the second stage for the actual localization. In this work we present novel, frequency domain, approaches for TDOA calculation in a reverberant and noisy environment. Our methods are based on the speech quasi-stationarity property, noise stationarity and on the fact that the speech and the noise are uncorrelated. The mathematical derivations in this work are followed by an extensive experimental study which involves static and tracking scenarios.

Keywords: Source localization; Non-stationarity; Decorrelation; TDOA

1. Introduction

Determining the spatial position of a speaker finds a growing interest in video conference scenarios where automated camera steering and tracking are required. Microphone arrays, which are usually used for speech enhancement in a noisy environment [22], can be used for the task of speaker localization as well [3,6,8,9,11,20,27]. The related algorithms can be divided into two groups: single and dual-step approaches. In single step approaches the source location is determined directly from the measured data (i.e. the received signals at the microphone array). In the dual-step approaches, the location estimate is obtained by applying two algorithmic stages. First, *time difference* (or *time delay*) *of arrival* (TDOA) estimates are obtained from different microphone

*Corresponding author. Tel.: +972 4 8294751; fax: +972 4 8292795.
E-mail addresses: dvorkind@tx.technion.ac.il (T.G. Dvorkind), gannot@eng.biu.ac.il (S. Gannot).
URL: http://www.eng.biu.ac.il/~gannot.

0165-1684/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2004.09.014

pairs. Then, these TDOA readings are used for determining the spatial position of the source.

Single step approaches can be further divided into two groups. The first group is the high-resolution spectral estimation methods. The well-known *multiple signal classification* (MUSIC) algorithm [35] is a member of this group. So is the work in [21] which considers direction of arrival (DOA) estimation with a uniform circular arrays that outperforms MUSIC-like algorithms at low *signal-to-noise ratio* (SNR) for similar computational loads. Though the mentioned algorithms can perform DOA estimation of multiple sources they are mainly suited for narrow-band signals. We note, however, that extension of those algorithms for a wide-band signals do exist. See for example [12,40]. In the second group of single step approaches we find the *maximum-likelihood* (ML) algorithms, which estimate the source locus by applying the ML criterion. Usually, the ML formulation leads to algorithms involving maximization of the output power of a beamformer steered to potential source locations (i.e. [3,6,8,9,11]).

In the dual-step approaches group, the first algorithmic stage involves TDOA estimation from spatially separated microphone pairs. The *maximum-likelihood generalized cross correlation* (ML-GCC)[1] method presented by Knapp and Carter [27] is considered to be the classical solution for this algorithmic stage. However, the GCC method assumes a reverberant-free model such that the *acoustical transfer function* (ATF), which relates the source and each of the microphones, is a pure delay. Champagne et al. showed this approximation to be inaccurate in reverberant conditions, which frequently occur in enclosed environments [7]. Consequently, algorithms for improving the GCC method in presence of room reverberation were suggested [5,38]. Unfortunately, the GCC method suffers from another model inaccuracy. It is assumed by the GCC model that the noise field is uncorrelated, an assumption which usually does not hold. Thus, the GCC method cannot distinguish between the speaker and a directional interference, as it tends to estimate the TDOA of the stronger signal. Directional interference usually occurs when a point source, e.g. computer fan, projector or a ceiling fan, exists. The authors in [31] suggested discriminating speaker from directional noise with a Gaussian mixture model. A different approach was presented in [10,30], where *higher order statistics* (HOS) was employed for TDOA estimation of a non-Gaussian source and correlated Gaussian noise.

Recently, subspace methods were suggested for TDOA estimation. Assuming spatially uncorrelated noise, Benesty suggested a time domain algorithm for estimating the (truncated to shorter length) impulse responses for TDOA extraction [2]. Extension of that work, for spatially correlated noise was presented by Doclo and Moonen [13,15]. Assuming that the noise correlation matrix is known (using a *voice activity detector* (VAD)), the authors presented a time domain algorithm for TDOA estimation using a *generalized eigenvalue decomposition* (GEVD) approach and a pre-whitening approach.

In this work, we tackle the TDOA estimation problem. Hence, the proposed solutions are members of the dual-step approaches. We address the TDOA extraction based on a single microphone pair. The second algorithmic stage, i.e. the actual localization based on multiple TDOA readings which are extracted from additional microphone pairs [4,18,25], is not addressed in this work. Our model assumptions consider reverberation and spatially correlated noise scenarios [17,19]. Specifically, we consider a single speaker in a stationary noise environment. In [22] the speaker's ATF-s ratio was used as part of a beamformer in a speech enhancement application. Here, we exploit this quantity for the source localization application. Particularly, we show that the TDOA reading can be extracted from the location of the maximal peak in the corresponding impulse response. Similar to [22] and the preceding work by Shalvi and Weinstein [37] we also assume that the interfering noise is relatively stationary, and present a framework where the ATF-s ratio and a noise related term are estimated simultaneously without any VAD employment. Quasi-stationarity of the speech and stationarity of the noise are exploited to derive batch and recursive

[1] For brevity we will simply notate this by GCC.