Robotics and Autonomous Systems 58 (2010) 1185–1196

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Speaker localization and tracking with a microphone array on a mobile robot using von Mises distribution and particle filtering

Ivan Marković*, Ivan Petrović

Faculty of Electrical Engineering and Computing, Department of Control and Computer Engineering, University of Zagreb, Zagreb, Croatia

ARTICLE INFO

Article history:
Received 13 April 2010
Received in revised form
26 July 2010
Accepted 3 August 2010
Available online 7 August 2010

Keywords:
Speaker localization
Microphone array
von Mises distribution
Particle filtering

ABSTRACT

This paper deals with the problem of localizing and tracking a moving speaker over the full range around the mobile robot. The problem is solved by taking advantage of the phase shift between signals received at spatially separated microphones. The proposed algorithm is based on estimating the time difference of arrival by maximizing the weighted cross-correlation function in order to determine the azimuth angle of the detected speaker. The cross-correlation is enhanced with an adaptive signal-to-noise estimation algorithm to make the azimuth estimation more robust in noisy surroundings. A post-processing technique is proposed in which each of these microphone-pair determined azimuths are further combined into a mixture of von Mises distributions, thus producing a practical probabilistic representation of the microphone array measurement. It is shown that this distribution is inherently multimodal and that the system at hand is non-linear. Therefore, particle filtering is applied for discrete representation of the distribution function. Furthermore, the two most common microphone array geometries are analysed and exhaustive experiments were conducted in order to qualitatively and quantitatively test the algorithm and compare the two geometries. Also, a voice activity detection algorithm based on the before-mentioned signal-to-noise estimator was implemented and incorporated into the existing speaker localization system. The results show that the algorithm can reliably and accurately localize and track a moving speaker.

1. Introduction

In biological lifeforms hearing, as one of the traditional five senses, elegantly supplements other senses as being omnidirectional, not limited by physical obstacles and the absence of light. Inspired by these unique properties, researchers strive towards endowing mobile robots with auditory systems to further enhance human–robot interaction, not only by means of communication but also, just as humans do, to make intelligent analysis of the surrounding environment. By providing speaker location to other mobile robot systems, like path planning, speech and speaker recognition, such a system would be a step forward in developing a fully functional human-aware mobile robot.

The auditory system must provide a robust and non-ambiguous estimate of the speaker location, and must be updated frequently in order to be useful in practical tracking applications. Furthermore, the estimator must be computationally non-demanding and possess a short processing latency to make it practical for real-time systems. The afore-mentioned requirements and the fact of an auditory system being placed on a mobile platform, thus having to respond to constantly changing acoustic conditions, make speaker localization and tracking a formidable problem.

Existing speaker localization strategies can be categorized in four general groups. The first group of algorithms refer to beamforming methods in which the array is steered to various locations of interest and searches for the peak in the output power [1–4]. The second group includes beamforming methods based upon analysis of a spatiospectral correlation matrix derived from the signals received at the microphones [5]. The third group relies on computational simulations of the physiologically known parts of the hearing system, e.g. binaural cue processing [6–8]. The fourth group of localization strategies is based on estimating the time difference of arrival (TDOA) of the speech signals relative to pairs of spatially separated microphones and then using that information to infer about the speaker location. Estimation of the TDOA and speaker localization from TDOA are two separate problems. The former is usually calculated by maximizing the weighted cross-correlation function [9], while the latter is commonly known as multilateration, i.e. hyperbolic positioning, which is a problem of calculating the source location by finding the intersection of at least two hyperbolae [10–13].

* Corresponding author. Tel.: +385 1 6129 561; fax: +385 1 6129 809.
E-mail addresses: ivan.markovic@fer.hr (I. Marković), ivan.petrovic@fer.hr (I. Petrović).
URLs: http://act.rasip.fer.hr/people-opis.php?id=199 (I. Marković), http://act.rasip.fer.hr/people-opis.php?id=1 (I. Petrović).

0921-8890/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.robot.2010.08.001

In mobile robotics, due to small microphone array dimensions, usually hyperbolae intersection is not calculated, only the angle (azimuth and/or elevation) is estimated [14–18].

Even though the TDOA estimation based methods are outperformed to a certain degree by several more elaborate methods [19–21], they still prove to be extremely effective due to their elegance and low computational costs. This paper proposes a new speaker localization and tracking method based on TDOA estimation, probabilistic measurement modelling based on von Mises distribution, and particle filtering. Speaker localization and tracking based on particle filtering was also used in [1,22–24], but the novelty of this paper is the proposed measurement model used for *a posteriori* inference about the speaker location. The benefits of the proposed approach are that it solves the front–back ambiguity, increases the robustness by using all the available measurements, and localizes and tracks a speaker over the full range around the mobile robot, while keeping low computational complexity of TDOA estimation based algorithms.

The rest of the paper is organized as follows. Section 2 describes the implemented azimuth estimation method and the voice activity detector. Section 3 analyses Y and square microphone array geometries, while Section 4 defines the framework for the particle filtering algorithm, introduces the von Mises distribution, the proposed measurement model, and describes in detail the implemented algorithm. Section 5 presents the conducted experiments. Finally, Section 6 concludes the paper and presents future works.

2. TDOA estimation

The main idea behind TDOA-based locators is a two-step one. Firstly, TDOA estimation of the speech signals relative to pairs of spatially separated microphones is performed. Secondly, this data is used to infer about speaker location. The TDOA estimation algorithm for two microphones is described first.

2.1. Principle of TDOA

A windowed frame of L samples is considered. In order to determine the delay $\Delta\tau_{ij}$ in the signal captured by two different microphones (i and j), it is necessary to define a coherence measure which will yield an explicit global peak at the correct delay. Cross-correlation is the most common choice, since we have at two spatially separated microphones (in an ideal homogeneous, dispersion-free and lossless scenario) two identical time-shifted signals. Cross-correlation is defined by the following expression:

$$R_{ij}(\Delta\tau) = \sum_{n=0}^{L-1} x_i[n]x_j[n-\Delta\tau], \tag{1}$$

where x_i and x_j are the signals received by microphone i and j, respectively. As stated earlier, R_{ij} is maximal when the correlation lag in samples, $\Delta\tau$, is equal to the delay between the two received signals.

The most appealing property of the cross-correlation is the ability to perform calculations in the frequency domain, thus significantly lowering the computational intensity of the algorithm. Since we are dealing with finite signal frames, we can only estimate the cross-correlation:

$$\hat{R}_{ij}(\Delta\tau) = \sum_{k=0}^{L-1} X_i(k)X_j^*(k)e^{j2\pi\frac{k\Delta\tau}{L}}, \tag{2}$$

where $X_i(k)$ and $X_j(k)$ are the discrete Fourier transforms (DFTs) of $x_i[n]$ and $x_j[n]$, and $(.)^*$ denotes complex-conjugate. We are windowing the frames with rectangular windows and no overlap. Therefore, before applying a Fourier transform to signals x_i and x_j, it is necessary to zero-pad them with at least L zeros, since we want to calculate linear, and not circular convolution.

A major limitation of the cross-correlation given by (2) is that the correlation between adjacent samples is high, which has an effect of wide cross-correlation peaks. Therefore, appropriate weighting should be used.

2.2. Spectral weighting

The problem of wide peaks in unweighted, i.e. generalized, cross-correlation (GCC) can be solved by whitening the spectrum of signals prior to computing the cross-correlation. The most common weighting function is the phase transform (PHAT) which, as has been shown in [9], under certain assumptions yields a maximum likelihood (ML) estimator. What the PHAT function ($\psi_{\text{PHAT}} = 1/|X_i(k)||X_j^*(k)|$) does, is that it whitens the cross-spectrum of signals x_i and x_j, thus giving a sharpened peak at the true delay. In the frequency domain, GCC-PHAT is computed as:

$$\hat{R}_{ij}^{\text{PHAT}}(\Delta\tau) = \sum_{k=0}^{L-1} \frac{X_i(k)X_j^*(k)}{|X_i(k)||X_j(k)|}e^{j2\pi\frac{k\Delta\tau}{L}}. \tag{3}$$

The main drawback of the GCC with PHAT weighting is that it equally weights all frequency bins regardless of the signal-to-noise ratio (SNR), thus making the system less robust to noise. To overcome this issue, as proposed in [1], a modified weighting function based on SNR is incorporated into the GCC framework.

Firstly, a gain function for such modification is introduced (this is simply a Wiener gain):

$$G_i^n(k) = \frac{\xi_i^n(k)}{1+\xi_i^n(k)}, \tag{4}$$

where $\xi_i^n(k)$ is the *a priori* SNR at the ith microphone, at time frame n, for frequency bin k and $\xi_i^0 = \xi_{min}$. The *a priori* SNR is defined as $\xi_i^n(k) = \lambda_{i,x}^n(k)/\lambda_i^n(k)$, where $\lambda_{i,x}^n(k)$ and $\lambda_i^n(k)$ are the speech and noise variance, respectively. It is calculated by using the *decision-directed estimation* approach proposed in [25]:

$$\xi_i^n(k) = \alpha_e[G_i^{n-1}(k)]^2\gamma_i^{n-1}(k) + (1-\alpha_e)\max\{\gamma_i^n(k)-1, 0\}, \tag{5}$$

where α_e is the adaptation rate, $\gamma_i^n = |X_i^n(k)|^2/\lambda_i^n(k)$ is the *a posteriori* SNR, and $\lambda_i^0(k) = |X_i^0(k)|^2$.

In stationary noise environments, the noise variance of each frequency bin is time invariant, i.e. $\lambda_i^n(k) = \lambda_i(k)$ for all n. But if the microphone array is placed on a mobile robot, most surely due to the robot's changing location, we will have to deal with non-stationary noise environments. An algorithm used to estimate $\lambda_i^n(k)$ is based on *minima controlled recursive averaging* (MCRA) developed in [26,27]. The noise spectrum is estimated by averaging past spectral power values, using a smoothing parameter that is adjusted by the speech presence probability. Speech absence in a given frame of a frequency bin is determined by the ratio between the local energy of the noisy signal and its minimum within a specified time window. The smaller the ratio in a given spectrum, the more probable the absence of speech is. Further improvement can be made in (4) by using a different spectral gain function [28].

To make the TDOA estimation more robust to reverberation, it is possible to modify the noise estimate $\lambda_i^n(k)$ to include a reverberation term $\lambda_{i,\text{rev}}^n(k)$:

$$\lambda_i^n(k) \mapsto \lambda_i^n(k) + \lambda_{i,\text{rev}}^n(k), \tag{6}$$

where $\lambda_{i,\text{rev}}^n$ is defined using a reverberation model with exponential decay [1]:

$$\lambda_{i,\text{rev}}^n(k) = \alpha_{\text{rev}}\lambda_{i,\text{rev}}^{n-1}(k) + (1-\alpha_{\text{rev}})\delta|G_i^{n-1}(k)X_i^{n-1}(k)|^2, \tag{7}$$