2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance

Sound Source Localization for Video Surveillance Camera

Jacek Stachurski, Lorin Netsch, Randy Cole

Texas Instruments
Embedded Signal Processing R&D Lab
Dallas, Texas

Abstract

While video analytics used in surveillance applications performs well in normal conditions, it may not work as accurately under adverse circumstances. Taking advantage of the complementary aspects of video and audio can lead to a more effective analytics framework resulting in increased system robustness. For example, sound scene analysis may indicate potential security risks outside field-of-view, pointing the camera in that direction. This paper presents a robust low-complexity method for two-microphone estimation of sound direction. While the source localization problem has been studied extensively, a reliable low-complexity solution remains elusive. The proposed direction estimation is based on the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) method. The novel aspects of our approach include band-selective processing and inter-frame filtering of the GCC-PHAT objective function prior to peak detection. The audio bandwidth, microphone spacing, angle resolution, processing delay and complexity can all be adjusted depending on the application requirements. The described algorithm can be used in a multi-microphone configuration for spatial sound localization by combining estimates from microphone pairs. It has been implemented as a real-time demo on a modified TI DM8127 IP camera. The default 16 kHz audio sampling frequency requires about 5 MIPS processing power in our fixed-point implementation. The test results show robust sound direction estimation under a variety of background noise conditions.

1. Introduction

Many video analytics solutions used in surveillance applications to identify security-risk events perform relatively well in normal conditions. They might not, however, detect emergency events accurately in cases of view obstruction, low or rapidly changing lighting, out-of-view activities, or other adverse conditions (e.g. rain, fog, smoke). In such cases, audio analytics can be used to provide additional information about the environment under surveillance. Audio analytics can analyze the sound scene of a surveyed environment and provide additional data about activities not readily discerned by a camera. Sound identification may alert to potential security risks and sound localization may be used to point the camera in the direction of interest. Taking advantage of the complementary aspects of video and audio can provide a powerful framework that should lead to increased system robustness and positive alarm detection rate.

This paper presents a prototype solution that adds sound localization capability to a surveillance camera. In addition to security, localization of a sound source is of interest in many other applications including video conferencing, smart buildings, robotics, and assisted living. In these intelligent-environment applications, microphone arrays are used to track speakers and various other sounds of interest. While the sound localization problem has been studied extensively, a reliable low-complexity solution remains elusive.

We generally take the human ability to localize sound for granted. Our auditory system combines multiple cues for an effective sound source localization including signal level differences between ears, temporal and spectral analyses, and pattern matching [1]. Effects such as the head shadowing of the sound, and direction-specific frequency patterns imposed by the outer ear and torso, help us localize the spatial sound origins. These are difficult to replicate using microphones and low complexity signal processing.

Sound source localization algorithms are traditionally based on the sound's Time Difference of Arrival (TDOA) between various microphones. One of the most widely used TDOA estimators employs Generalized Cross-Correlation with Phase Transform (GCC-PHAT) [2, 3]. The GCC-PHAT method is attractive because, while being sub-optimal under ideal conditions, it tends to perform well in challenging environments, for example in the presence of reverberation. Several recent approaches aimed at improving accuracy of the TDOA estimate modify the GCC-PHAT weighting function, for example by applying an SNR-dependent exponent to the weighting function [4], adding a bias term in the denominator [5],

978-1-4799-0703-8/13/$31.00 ©2013 IEEE

and using estimates of the phase statistics [5, 6]. Other approaches extend GCC-PHAT by calculating the TDOA in temporal and frequency bands followed by merging the estimates [7, 8], or including TDOA prior assumptions during estimation [10]. To reduce the effects of noise, some methods pre-process the input to remove unwanted signal components prior to applying GCC-PHAT, for example by performing spectral subtraction and mean normalization [5], or by decomposing the input using basis functions [9]. The GCC-PHAT output can be post-processed to obtain higher TDOA estimate accuracy, such as applying interpolation functions during the search for the GCC-PHAT delay peak value [4].

All the above approaches improve the TDOA accuracy but not necessarily its stability. Real-time applications require not only an accurate, but also a stable and low-latency TDOA estimate in order to accomplish objectives such as focusing a camera in the direction of a sound without annoying camera jitter. In such applications, estimation of the precise location of the sound source is often not as important as ensuring the stability of the estimate. Stable TDOA estimation often implies high sampling rates and processing large windows of data, which increases latency and consumes processing resources. In this paper, we present a low-resource method based on GCC-PHAT to provide accurate and stable sound-direction estimates. TDOA estimates are known to exhibit anomalous behavior during short periods of silence and in noise [11]. In our approach, we select only the frequency bands that may most likely contain useful information, and reject those dominated by noise. We propose a novel inter-frame adaptive filter which is applied to the GCC-PHAT objective function prior to determining the TDOA estimate. The inter-frame filter reinforces correlation peaks across a number of processing frames and thus substantially improves stability of the final estimate. The inter-frame filtering is performed before a peak is identified in the GCC-PHAT objective function, effectively smoothing the direction estimate. To further improve performance, post-processing median-filter and hysteresis are added.

We implemented the described sound-direction algorithm on the TI DM8127 DaVinci processor in the Appro-built TMDSIPCAM8127J3 IP camera. The camera reference design incorporates only one built-in microphone which is not sufficient for our purpose; we disabled this built-in microphone, and connected two external microphones to the AIC3204 audio codec in the IP camera. The added connections provide balanced inputs and microphone bias. The two external microphones are attached to moveable mounts located directly in front of the camera to test performance at various microphone separation distances. To turn the camera in the sound direction, we mounted it on a high-torque servo motor. We use the AIC3204 codec's audio output to generate the PWM drive signal required by the servo. Video output from the camera is available from HDMI or by streaming through the internet.

We describe the proposed sound direction algorithm in Section 2 and the modified IP camera and its steering system in Section 3, followed by a discussion of the tests and results in Section 4.

2. Estimation of Sound Direction

The block diagram of our two-microphone sound direction estimator is shown in Fig. 1. The processing blocks are organized into three categories: front-end processing, direction estimator, and post-processing.

2.1. Front-end processing

The audio input is split into overlapping data blocks and a Hamming window is applied to each block, followed by the FFT. The offset between the successive data blocks is

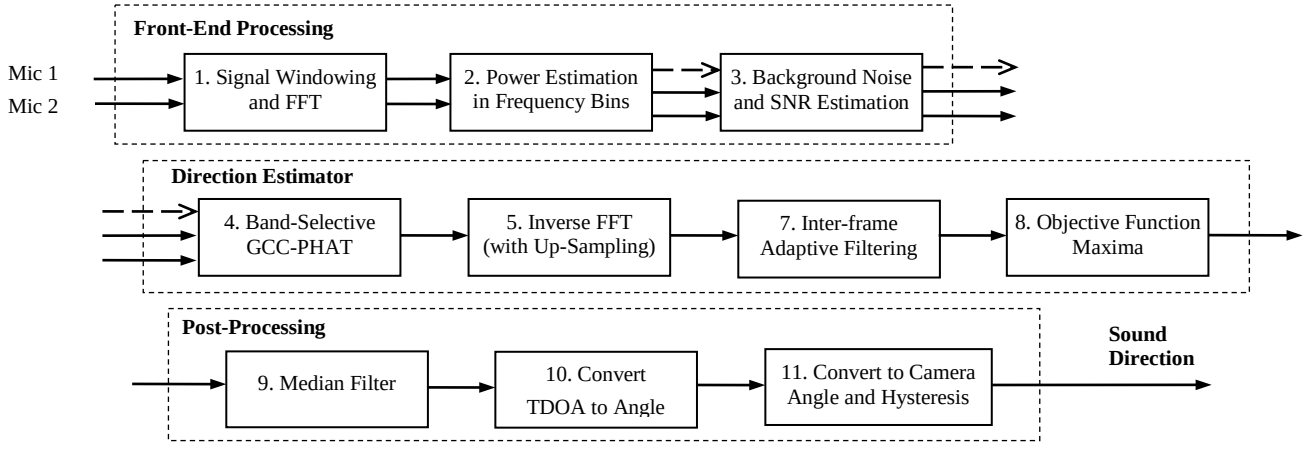

Figure 1: Block diagram of the two-microphone sound direction estimator

referred to as a processing "frame". One FFT is performed for each of the two data channels, that is two FFTs for each processing frame.

The signal power is estimated in each frequency band and overall, and SNR is calculated based on the estimated background noise. Using thirteen frequency bands for 16 kHz input signal, for example, we calculate fourteen SNR values per frame (one per each frequency band, and one over all frequencies). The SNR estimates in each frequency band are used for band selection: frequencies that likely contain useful information are included in the GCC-PHAT analysis, while others (those dominated by noise) are not; we use thresholds of 0-10 dB depending on the desired sensitivity. In addition, if the overall SNR of a particular frame is below a specified threshold, e.g. 10-15 dB, the frame is not used for TDOA estimation.

2.2. Direction estimator

To improve the performance of a baseline GCC-PHAT, we apply spectral band selection based on the estimated SNR. Frequency selection addresses the weakness of GCC-PHAT which uses only the signal phase in its TDOA estimation. All the phase values contribute to the TDOA estimate regardless of their substance; that is, the potentially random phase of noise is treated at par with the signal of interest phase. By selecting the spectral bands dominated by the sound of interest and discarding those dominated by noise, the final TDOA estimate is improved.

The SNR-based band selection and the PHAT weighing (based on spectral amplitude) are combined

$$W_k = B_k \, W_{k\,\mathrm{PHAT}} \quad \text{with} \quad W_{k\,\mathrm{PHAT}} = \left| X_k Y_k^* \right|^{-1} \tag{1}$$

and the GCC-PHAT values at each frequency k are calculated as

$$G_k = W_k X_k Y_k^* \tag{2}$$

where X and Y represent FFTs of the two input channels.

Up-sampling is performed in the frequency domain by inserting zeros in the spectral representation before the inverse transform. This interpolation increases the resolution of the TDOA estimate, and thus the resolution of the sound direction estimate.

An inverse FFT of G_k is performed to obtain the TDOA objective function $g(d)$. The offset d of the maximum value of $g(d)$ would provide the TDOA estimate in the basic GCC-PHAT. We include, however, additional inter-frame adaptive filtering prior to this maximum search such that

$$\tilde{g}(d) = \sum_{m=1}^{M} a_m g_m(d) \tag{3}$$

where M is the filter size and a_m are the adaptive filter coefficients.

The inter-frame filtering is applied to smooth out the frame-to-frame maxima variations in the objective functions $g_m(d)$ domain. The filtering is applied across corresponding elements of $g(d)$ from multiple frames (which deemphasizes temporal maxima), not within each $g(d)$ function (which could smear the maxima). This inter-frame filtering reinforces correlation peaks across a number of processing frames and thus substantially improves stability of the final estimate. While various filters may be employed, we found that setting the filter coefficients a_m to the frame's signal power works sufficiently well. This adaptive filtering may also be viewed as amplitude-weighted average of the TDOA objective functions $g(d)$ across frames. The advantages of such filtering are low-delay tracking of high-energy sound onsets and very low complexity (since only a limited number of the $g(d)$ objective function values need to be considered). For example, with the 344 m/s speed of sound, 10 cm microphone distance, 16 kHz input sampling frequency, and the up-sampling rate of 4, the maximum TDOA values are ± 9 and so only 19 points of $g(d)$ are filtered.

The offset d of the maximum of the filtered objective function $\tilde{g}(d)$ provides our initial TDOA estimate as

$$\mathrm{TDOA} = \arg\max_d \tilde{g}(d) \tag{4}$$

The estimated TDOA is then further smoothed out in the post-processing stage.

2.3. Post-processing

In post-processing, we use a median filter to smooth out occasional spikes in the TDOA estimates. We use a 5-tap filter; a longer filter may be used if the delay is acceptable. Other smoothing filters also may be applied. Hysteresis is further used to prevent a frame-to-frame jitter between adjacent TDOA values. The TDOA estimate is kept unchanged from the last frame if it varies by less than a specified value from the previous estimate; otherwise the new estimate is accepted.

The TDOA estimate is converted to an angle specifying the sound direction. The standard far-field assumption of the sound source is used to calculate the angle of sound arrival.

3. IP Camera with Sound Localization

We implemented the sound source localization algorithm on the TI DM8127 IP camera. It runs on the ARM Cortex A8 processor of the DM8127 chip under the Linux OS provided with the camera. The C674x DSP (and the on-board video coprocessor) performs video analysis. We use the AIC3204 codec's audio output to generate the PWM drive signal required by the servo that turns the camera towards the identified sound.

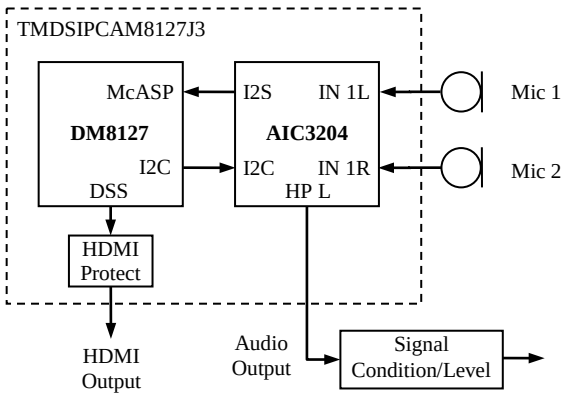

Figure 2: Modified IP Camera

3.1. Microphone configuration

The DM8127 IP camera includes a single built-in microphone connected to the AIC3204 codec's MIC1R P/M balanced inputs. Blocking capacitors on the sub board isolate the codec inputs from the microphone, and the codec's MICBIAS output provides bias to the microphone through a 1K ohm resistor.

Our prototype system, shown in Fig. 2, requires two microphone signals to perform sound localization. To achieve that, we disconnected the built-in microphone and attached the MIC1R P/M inputs to a connector on the side of the camera. We duplicated the blocking capacitor and microphone bias circuitry for the codec's MIC1L P/M inputs and attached them to a second connector. We used two small external omni-directional microphones mounted on an adjustable camera base to test several microphone configurations. In order to use the two microphones, we modified the ALSA audio driver software in the ARM Linux distribution supplied with the camera which had originally been set for single-microphone use. We used ALSA utilities to configure the AIC3204 codec for the required input gains and pre-filter characteristics.

3.2. Camera rotation

To enable the camera's rotation towards the direction of a detected sound, we mounted it on an HS-985MG high-torque servo. We utilized the AIC3204 codec HP output to produce the PWM drive signal. We sent the codec PWM signal through an external signal conditioning circuit that provided hysteresis for noise stability, gain, level shifting, and fast rise-time PWM pulse shaping to sharpen signal transition edges required by the servo.

We developed a software module to generate the continuous servo control audio PWM signal in a separate running thread. The module adjusts the PWM signal width to rotate the camera to the desired position. A sampling frequency of 16 kHz yields 6.25 ms PWM pulse width which equates to an angular resolution of 6.25 degrees. For our experiments, this was adequate to position the camera so the located sound source is well within its field of view.

4. Tests and Results

The default settings of our sound direction estimator include 10 cm microphones spacing, 16 kHz audio sampling frequency, 1024-point FFT, and 16 Hz analysis rate (1000 samples window shift). The 8x frequency-domain up-sampling increases the effective sampling frequency to 128 kHz. With the sound speed of 343 m/s, these settings provide an angle resolution of about 3 degrees near the center, and 15 degrees near the ±90 angles. The far-field assumption [12] requires at least 47 cm distance between sound and microphones.

We split the bandwidth into thirteen frequency bands from 100 to 7000 Hz (with the bands varying from 200 Hz to 1000 Hz at the lower and upper ends of the spectrum, respectively). Only bands with the SNR of 5 dB or higher are used in the estimation. The inter-frame adaptive filter coefficients are set to the signal power of eight most recent frames. The size of the median post-filter is five. These default settings at 16 kHz audio sampling frequency take about 5 MIPS processing power in our fixed-point implementation.

All of the above parameters can be set as needed. We tested various combinations and validated consistent performance across a wide range of settings. A desired combination should be selected based on the required angle resolution, microphone spacing, noise environment, as well as computational complexity and delay constraints.

A performance example of the algorithm with default settings is presented in Fig. 3. The example represents a 25 second recording of a moving speaker in fairly quiet office environment. As the speaker moves from the center to the left and then to the right, we desire a steady sound-direction estimate that tracks the sound location. The speech signal (mono down mix) is shown in Fig. 3a. In Fig. 3b, the estimate obtained from the GCC-PHAT baseline algorithm is shown. This estimate is very noisy with many excursions from the correct sound direction. The plot well illustrates one of the weaknesses of the basic GCC-PHAT method: as the amplitude spectrum is normalized through the PHAT transformation, frequencies representing the signal of interest and those with mostly noise get similar consideration, often resulting in a noisy outcome. Fig. 3c shows the baseline GCC-PHAT with post-processing applied. It can be observed that a number of excursions from the sound path are being eliminated and the direction estimate becomes much "smoother". Ever more sophisticated post-processing could further clean up this estimate at the expense of computational complexity and/or added delay. Finally, Fig. 3d presents

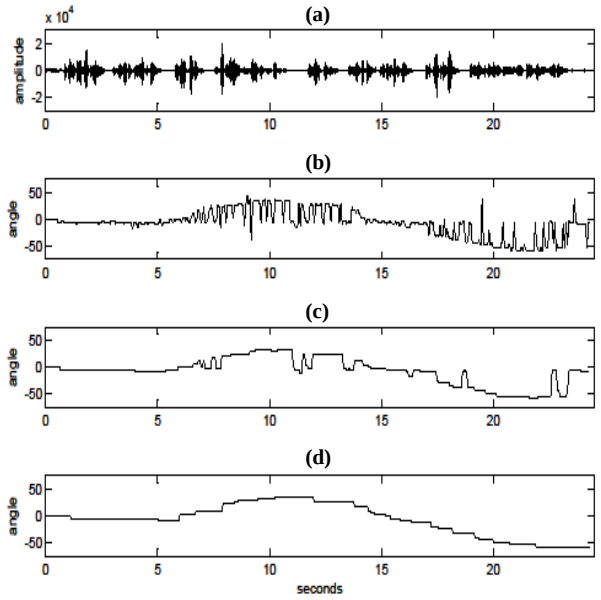

Figure 3: (a) Speech example, (b) GCC-PHAT estimate, (c) GCC-PHAT with post-filtering, (d) Proposed algorithm's output

the output of the proposed sound direction algorithm. Selection of the higher SNR frequency bands and, more importantly, the inter-frame adaptive filtering of the objective function smoothes out the undesired frame-to-frame GCC-PHAT variations and provides a steadier sound-direction estimate.

We performed a set of tests to determine the expected performance of the proposed sound-direction algorithm in clean and noisy conditions. While the quantitative tests were performed with speech signals, no part of the described algorithm assumes speech-like signal characteristics and we have also verified that directions of other sounds are equally well identified. Five female and five male speakers were recorded while slowly moving around the microphones (left and right, closer and further away). Four background noises were also recorded with the same microphone configuration. To simulate a defused character of the noise, multiple loudspeakers were used facing away from the microphones. Babble, car, office, and street noises were recorded. The background noise was then mixed with the speech at SNR levels from 25 to 5 dB. For the clean condition, we calculated the average value of frame-to-frame angle change which we found to be a good simple performance indicator for a stationary or slowly moving sound source. When the estimator matches the slowly-moving sound direction well, the average angle change is low; as the algorithm's performance deteriorates, more and more excursions from the correct values increase this average. Note that the average angle change heavily depends on the test material so it is not suitable to be an objective performance measure in itself. It is, however, a helpful indicator that under given test conditions one

	baseline GCC-PHAT	with pp (post-processing)	with pp & inter-frame filtering	with pp & band-selection	with all
Avg	142.1	30.4	15.1	5.1	2.8
Std	13.5	4.6	3.6	1.4	0.9

Table 1 Average per second frame-to-frame angle change

SNR	babble	car	office	street
25	0.2	0.1	0.2	0.2
15	0.9	0.6	0.8	0.7
5	3.0	2.0	4.2	5.2

Table 2 Average per-frame angle difference, clean vs. noise

estimator performs better than another, and it indicates well when an algorithm may be starting to break down. The average per-second frame-to-frame angle change and its standard deviation for the baseline GCC-PHAT, with post-processing, with band-selection, with inter-frame filtering, and with all of the above are summarized in Table 1. The 2.8°/sec average angle change of our algorithm provides a much more stable outcome than the reference 142.1°/sec baseline GCC-PHAT estimate. The large average angle change of GCC-PHAT is caused by the frequent excursions from the correct sound direction as can be observed in Fig. 3a.

For the noisy conditions, we calculated the average per-frame angle difference between the clean speech direction and the corresponding estimates with background noise added. These tests are designed to validate the proposed algorithm stability even in environments with considerable noise. In the tests, the average angle difference between clean and noisy speech varied between 0.1°/frame for 25dB SNR up to 5.2°/frame for 5dB SNR signal as presented in Table 2. For reference, the average per-frame angle difference between the proposed algorithm and the baseline GCC-PHAT for the tested database was 15°/frame (much higher than 5.2°/frame). Understandably, the difference increases with increased noise level. Most of the time, the angle estimates for clean and noisy speech are in fact the same – the difference mostly occurs in timing of the transitions between angles as a speaker moves around. The results do indicate that the proposed algorithm maintains stable performance even in considerable background noise.

The presented sound direction algorithm was incorporated into a real-time demo running on a modified DM8127 IP camera which is steered towards an estimated sound source. In this application, estimation of the precise angle of the sound source is not crucial (the sound source needs to appear in the camera's field of view, but it does not necessarily have to be centered). Ensuring the stability of the estimate to prevent unnecessary camera jitter is

more important. The demo was tested in a number of practical scenarios. One of the observations from the tests is that sound reflections may cause an incorrect detection of the direction angle, particularly in small office spaces when a speaker is facing away from the microphones. It may not be possible to totally eliminate this problem as the dominant sound that is recorded by the microphones does in fact arrive from a different location (a reflecting surface) than it originated from. As such, an estimator may correctly identify the dominant sound direction which, however, does not in this case coincide with the position of the original sound source. A more elaborate sound localization tracker would be needed in such situations.

5. Conclusions

Taking advantage of the complementary aspects of video and audio can lead to a powerful analytics framework that improves surveillance effectiveness and robustness. Sound identification may alert to potential security risks even when they are obstructed, hidden, or before they appear within camera field-of-view, and sound localization may point the camera in the direction of interest. We presented a robust low-complexity method for two-microphone estimation of sound direction based on TDOA with GCC-PHAT. Audio bandwidth, microphone spacing, angle resolution, processing delay and complexity can all be adjusted depending on the application requirements. The described algorithm can be used in a multi-microphone configuration for spatial sound localization by combining estimates from microphone pairs. The low-complexity techniques applied to better the sound direction estimate include spectral band selection and inter-frame adaptive filtering applied to the GCC-PHAT objective function. Spectral selection takes advantage of the frequency bands that most likely contain useful information, and rejects those dominated by noise. The novel inter-frame filtering of the GCC-PHAT objective function reinforces correlation peaks across a number of processing frames and thus substantially improves stability of the final estimate. This inter-frame filtering is performed before a correlation peak is identified, effectively smoothing the sound direction estimate. The above techniques may also be used with other time or frequency-domain TDOA estimators, for example with the simpler GCC estimator, to further reduce the system complexity. The results show significant improvement over the baseline GCC-PHAT method in terms of the output stability. The tests confirm a robust performance for clean speech as well as for a variety of background noise conditions. No part of the described algorithm assumes speech-like signal characteristics and the direction estimation performs equally well for any sound of sufficient SNR with respect to the estimated background noise. The proposed algorithm has been implemented as a real-time demo on a modified DM8127 IP camera in which two external microphones and a high-torque servo motor are added. The default 16 kHz audio sampling frequency requires only about 5 MIPS processing power in our fixed-point implementation.

6. References

[1] Woodruff, J., DeLiang Wang, "Binaural Localization of Multiple Sources in Reverberant and Noisy Environments", Audio, Speech, and Language Processing, IEEE Transactions on, On page(s): 1503 - 1512 Volume: 20, Issue: 5, July 2012

[2] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-24(4), pp.320-327, Aug. 1976

[3] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 375–378, 1997

[4] Bo Qin, Heng Zhang, Qiang Fu, Yonghong Yan, "Subsample Time Delay Estimation via Improved GCC PHAT Algorithm", Proc. ICSP 2008, pp. 2979-2982, 2008

[5] Hong Liu and Miao Shen, "Continuous Sound Source Localization based on Microphone Array for Mobile Robots", IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4339-4339, 2010

[6] Bowon Lee, Amir Said, Ton Kalker, and Ronald W. Schafer, "Maximum Likelihood Time Delay Estimation with Phase Domain Analysis in the Generalized Cross Correlation Framework", Workshop on Hands-free Speech Communication and Microphone Arrays, pp. 89-92, 2008

[7] Shoko Araki, Masakiyo Fujimoto, Kentaro Ishizuka, Hiroshi Sawada, and Shoji Makino, "A DOA Based Speaker Diarization System For Real Meetings", Workshop on Hands-free Speech Communication and Microphone Arrays, pp. 29-32, 2008

[8] Heidi Christensen, Ning Ma, Stuart N. Wrigley, Jon Barker, "A Speech Fragment Approach To Localising Multiple Speakers In Reverberant Environments", Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4593-4596, 2009

[9] Xiao Wu,Shijiu Jin, Zhoumo Zeng, Yunkui Xiao, Yajuan Cao, "Location for Audio signals Based on Empirical Mode Decomposition", Proc. of the IEEE International Conference on Automation and Logistics Shenyang, China, pp. 1888-1891, August 2009

[10] Bowon Lee and Ton Kalker, "Maximum a Posteriori Estimation of Time Delay", International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, pp. 285-288, 2007

[11] Anthony Badali, Jean-Marc Valin, Francois Michaud, and Parham Aarabi, "Evaluating Real-time Audio Localization Algorithms for Artificial audition in Robotics", IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2033-2038, 2009

[12] Ali Pourmohammad and Seyed Mohammad Ahadi, "Real Time High Accuracy 3-D PHAT-Based Sound Source Localization Using a Simple 4-Microphone Arrangement", IEEE Systems Journal, vol. 6, no. 3, pp. 455-468, Sep. 2012