CONSIDERING THE SECOND PEAK IN THE GCC FUNCTION FOR MULTI-SOURCE TDOA ESTIMATION WITH A MICROPHONE ARRAY

Dirk Bechler and Kristian Kroschel

Institut für Nachrichtentechnik
Universität Karlsruhe
Kaiserstr. 12, 76128 Karlsruhe, Germany
{bechler,kroschel}@int.uni-karlsruhe.de

ABSTRACT

Time Difference Of Arrival (TDOA) estimates can be used for passive acoustic multiple sound source localization with microphone arrays. The TDOA estimation is based on the cross-correlation function of two signals of a microphone pair. Existing systems assume only one dominant sound source per analysis frame rejecting the localization information of other active sources present in the acoustic environment. In this publication the possibility of simultaneously considering the localization information of two sound sources per analysis frame is studied. To measure the reliability of the information for the second active sound source, two properties of the cross-correlation function can be used. Real data experiments are carried out for a stationary two-speaker scenario in a noisy and reverberant office room.

1. INTRODUCTION

The need for acoustic perception of a humanoid robot is not only indispensable for man-machine interaction but also for the robot's capability of autonomously analyzing the acoustic scene. This acoustic scene analysis comprises the localization, separation and classification of sound sources present in the sound environment of the robot. Thereby, the localization plays the major role as the information of the positions of the sound sources can be used to facilitate the separation and classification task.

The most common technique for single sound source localization by means of a microphone array is based on the estimation of Time Delays Of Arrival (TDOAs) in an analysis frame in the microphone pairs of the sensor array. These TDOAs are commonly determined with the Generalized Cross-Correlation (GCC) method [1]. To have intersample precision, the mean value of the TDOA estimations of several successive analysis frames in the different microphone pairs is taken together with the knowledge about the array geometry to localize the sound source in 3D. To avoid the computational demanding solution of a set of nonlinear equations for the exact sound source position, suboptimal closed-form location estimators with sufficient precision [2, 3] can be used. In a multi-source environment, the existing TDOA estimators assume one dominant source in an analysis frame, delivering only one estimate per frame and rejecting the localization information of the second one.

In this publication, the possibility of delivering information of the simultaneous localization of two sound sources per analysis frame is investigated for a two-speaker scenario in a noisy and reverberant office environment. The additional information on the second sound source is of great interest, as the closed-form localization algorithm needs accurate TDOA estimates guaranteed by the averaging of successive estimates. Using the localization information of only one sound source per analysis frame reduces the number of estimates available for averaging for each of the different active sound sources leading to a loss of estimation precision and hence a loss in localization performance. Due to noise and reverberation influences complicating the TDOA estimation and due to the fact that in analysis frames the second sound source has often low energetic influence, the estimation of the second peak risks to be not very reliable desiring a confidence measure of the TDOA estimates of the second sound source. Therefore, two possible confidence features giving information about the reliability of the estimates of the second sound source are studied: the value of the 2^{nd} peak and the value of the ratio between the 2^{nd} and the 3^{rd} peak of the GCC function.

2. SINGLE SOURCE SCENARIO

2.1. Signal Model

For a given pair of spatially separated microphones M_i and M_j, the recorded sensor signals $x_i(t)$ and $x_j(t)$ for a signal $s(t)$, emanated from a remote sound source in a reverberant and noisy environment, can be modeled mathematically as

$$\begin{aligned} x_i(t) &= h_i(t) * s(t) + n_i(t) \\ x_j(t) &= h_j(t) * s(t-\tau_{ij}) + n_j(t), \end{aligned} \qquad (1)$$

where τ_{ij} represents the relative signal delay of interest, $*$ signifies the convolution operator, $h_i(t)$ is the acoustic impulse response between the sound source and the i^{th} microphone and the additive term $n_i(t)$ summarizes the channel noise in the microphone system as well as the environmental noise for the i^{th} sensor. This noise $n_i(t)$ is assumed to be uncorrelated with $s(t)$ and $n_j(t)$.

2.2. TDOA Estimation with GCC Method

The most popular approach for determining the TDOAs is called the Generalized Cross-Correlation method [1]. The relative time delay τ_{ij} is estimated as the time lag with the global maximum peak in the GCC function $R_{ij}^{(g)}(\tau)$

$$\hat{\tau}_{ij} = \underset{\tau}{\text{argmax}}\, R_{ij}^{(g)}(\tau). \quad (2)$$

The GCC function $R_{ij}^{(g)}(\tau)$ is defined as

$$R_{ij}^{(g)}(\tau) = \int_{-\infty}^{+\infty} \psi_{ij}(\omega) X_i(\omega) X_j(\omega)^* e^{j\omega\tau} d\omega. \quad (3)$$

The weighting function $\psi_{ij}(\omega)$ intends to decrease influences of noise and reverberation and tries to emphasize the GCC value at the true TDOA value τ_{ij}. For real environments the Phase Transform (PHAT) technique [1] has shown best performance. This PHAT weighting function is defined as

$$\psi_{ij}^{PHAT}(\omega) = \frac{1}{|X_i(\omega)X_j(\omega)^*|}. \quad (4)$$

3. MULTI-SOURCE SCENARIO

Neglecting room reverberation and noise influences, the signal model of a two sound source scenario can be written as follows:

$$\begin{array}{rcl} x_i(t) & = & \alpha_i s_1(t-\tau_{i1}) + \beta_i s_2(t-\tau_{i2}) \\ x_j(t) & = & \alpha_j s_1(t-\tau_{i2}) + \beta_j s_2(t-\tau_{j2}), \end{array} \quad (5)$$

where α_i and β_i are distance attenuation factors and τ_{i1} and τ_{i2} the TDOAs of sound source signal $s_1(t)$ and $s_2(t)$ for the i^{th} sensor. For this signal model, the cross-power spectrum $G_{ij}(\omega) = X_i(\omega)X_j(\omega)^*$ can be rewritten as follows:

$$\begin{array}{rcl} G_{ij}(\omega) & = & \alpha_i\alpha_j \mid S_1(\omega) \mid^2 e^{-j\omega(\tau_{j1}-\tau_{i1})} \\ & & +\beta_i\beta_j \mid S_2(\omega) \mid^2 e^{-j\omega(\tau_{j2}-\tau_{i2})} \\ & & +\alpha_i\beta_j S_1(\omega)S_2(\omega)^* e^{-j\omega(\tau_{j2}-\tau_{i1})} \\ & & +\alpha_j\beta_i S_1(\omega)^* S_2(\omega) e^{-j\omega(\tau_{j1}-\tau_{i2})}. \end{array} \quad (6)$$

Only if $s_1(t)$ and $s_2(t)$ are uncorrelated, the cross-correlation terms in the last two lines of formula (6) become zero and the GCC method can accurately estimate the TDOA values of the two sound sources with the values $\tau_{ij}^{(1)} = \tau_{j1} - \tau_{i1}$ for source 1 and $\tau_{ij}^{(2)} = \tau_{j2} - \tau_{i2}$ for source 2. Otherwise, peaks at incorrect TDOA values could be estimated. The number of wrong peaks in the GCC function will increase even more in real environments due to noise and reverberation influence.

Existing TDOA-based methods for multi-sound source localization using the GCC function take merely the strongest peak in the GCC function into account assuming that in most frames only one source contributes predominant energy and the associated TDOA estimate is a valid representation of that source's true TDOA [4, 5]. Another reason for only considering the strongest peak is the difficulty in determining if two sound sources are simultaneously active with sufficient energy to deliver two predominant peaks. As described above, noise and reverberation influences lead to erroneous peaks rendering the second peak more unreliable compared with the first one. Therefore, information about the reliability of the second peak by means of confidence criteria is desired.

4. RELIABILITY CRITERIA

For a single source scenario [6] and for a multi-source scenario considering just the first peak in the GCC function [7], two confidence criteria can be used very efficiently to evaluate the reliability of the actual TDOA estimate: the absolute value of the maximum peak and the ratio between the 1^{st} and the 2^{nd} peak in the GCC function. These criteria allow a reliability scoring of individual estimates and can be used to reject erroneous measurements. The higher the value of these properties of the GCC function is, the higher is the probability that the TDOA was estimated correctly. For highest values of these reliability criteria, a correct TDOA estimation of over 96% can be achieved [6, 7].

Assuming merely one predominant source in a two-speaker scenario gives information about the localization of only one sound source per analysis frame, rejecting the information about the second active speaker's position. This paper studies the possibility to take into account the information provided by the second sound source and to deliver TDOA estimates of both sound sources per analysis frame. For this study, the influence of the value of the second peak in the GCC function on the percentage of correct TDOA estimates for the second sound source is analyzed. As for the first peak, we wish that the higher the value of the second peak is, the higher is the probability that the TDOA estimate corresponding to the second sound source is correct. Likewise we wish that the higher the 2^{nd} peak dominates the 3^{rd} one, the higher is the probability of a confident TDOA measurement for the second sound source.

5. EXPERIMENTS

For data recording, a microphone array of 5 omni-directional electret condenser microphones in an equilateral double-tetrahedron geometry with a side length of $D = 28$ cm was used (Fig. 1). To evaluate the confidence criteria, real experiments were carried out in an office room of 5 m x 5 m x 3 m. To evaluate the performance of the TDOA estimator if multiple sound sources are present at the same time, two sets of recordings were made. In the first set (*Double-Talk*) two

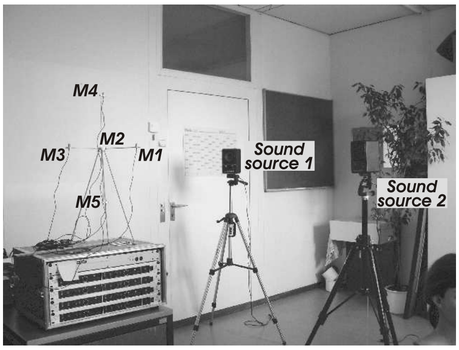

Fig. 1. Experimental setup

speakers utter different German sentences simultaneously. In the second set (*Extreme Double-Talk*) two speakers utter exactly identical German sentences synchronously. The speech signals were pre-recorded and played back by two loudspeakers. These two double-talk scenarios ensure significant periods of signal overlap of the two active sound sources. Combinations of female-female, male-female and male-male speakers were used for recordings (altogether 1012 words per speaker). The two loudspeakers were placed in 13 different positions $\{S_1S_2, S_1S_3, \ldots, S_1S_{10}, S_6S_7, S_6S_8, S_6S_9$ and $S_6S_{10}\}$ in the office room with typical environmental noise (fans, mechanical equipment, ...) and relatively strong reverberations (reverberation time $\mathrm{T}_{60} = 360$ ms). The height of the reference microphone M_1 and the sound sources was 1.5 m. For the x- and y-coordinates of the sound source positions see Fig. 2. The sampling frequency was $f_s = 16$ kHz. The recorded speech signals were analyzed in frames of 32 ms to assure quasi-stationarity. For this data segmentation a Hamming window with a 50% overlap was applied. In order to exclude frames containing silence, a simple energy threshold was used which guarantees that the TDOA estimation is accomplished only during speech activity. A TDOA estimation in the microphone pair M_iM_j is deemed correct if the product of the sampling frequency f_s and the term $|\tau_{ij}^{est} - \tau_{ij}^{\{1,2\}}|$, i.e. the absolute value of the difference of the estimated and the real TDOA value of sound source 1 *or* 2, is less than a decision threshold of $\mathrm{T_{dec}} = 1.5$ samples

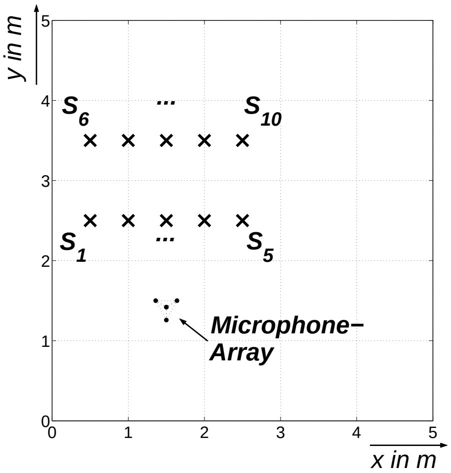

Fig. 2. Microphone and source positions

$$f_s \cdot |\tau_{ij}^{est} - \tau_{ij}^{\{1,2\}}| \begin{cases} \leq \mathrm{T_{dec}} & : \text{ correct} \\ > \mathrm{T_{dec}} & : \text{ false.} \end{cases} \tag{7}$$

6. RESULTS

In this section the results for the reliability criteria as described in Sect. 4 and for the two-speaker scenario as specified in Sect. 5 are presented. For this investigation, the TDOA estimates were divided into 8 intervals with increasing values for each of the two corresponding criteria: the maximum value of the second peak and the ratio between the second and the third peak in the GCC function of an analysis frame. The interval borders were chosen such that the number of TDOA estimates per interval is similar for all 8 intervals.

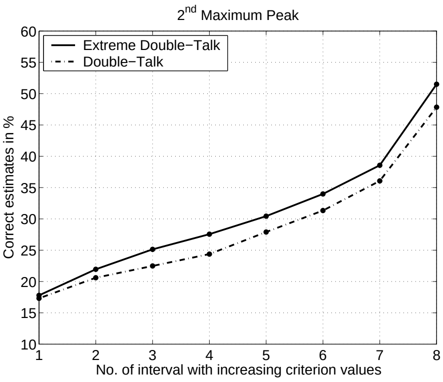

Fig. 3. Percentage of correct estimates for the confidence criterion of the 2^{nd} maximum peak

Fig. 3 illustrates the percentage of correct TDOA estimates as a function of the value of the second largest peak in the GCC function for the *Double-Talk* and the *Extreme Double-Talk* scenario. As expected, the percentage of correct TDOA estimates rises strongly with increasing values of the second peak. With an important percentage increase of 30.51% [33.72%] for the *Double-Talk* [*Extreme Double-Talk*] scenario between frames including highest (interval 8) and lowest (interval 1) second maximum peak values, this criteria shows good capabilities for reliability scoring of the TDOA estimates of the second sound source.

Compared with the criterion of the 2^{nd} peak, the confidence criterion of the ratio between the 2^{nd} and 3^{rd} peak of the GCC function is less appropriate to measure the reliability of the TDOA estimate of the second sound source. Though the percentage of correct estimates rises for increasing values of this ratio (Fig. 4), the percentage spread between interval 8 and 1 of 15.43% [20.58%] indicates by

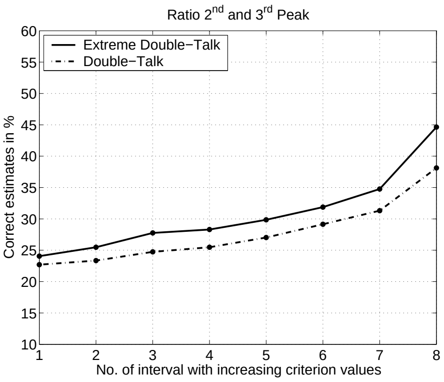

Fig. 4. Percentage of correct estimates for the confidence criterion of the ratio between 2^{nd} and 3^{rd} peak

comparison a more limited applicability as a reliability indicator.

The better performance of the criteria for the *Extreme Double-Talk* scenario reaching higher percentages of correct TDOA estimates compared to the *Double-Talk* scenario can be explained by the fact that for the *Extreme Double-Talk* scenario the signal overlap of the two sound sources is more important and due to the synchrony of the utterances both sound sources contribute with equal signal power leading to two peaks in comparable order of magnitude in the GCC function.

With 47.83% [51.51%] of correct estimates for highest values of the 2^{nd} peak criterion (Fig. 3), and with 38.13% [44.64%] for highest values of the criterion of the ratio between the 2^{nd} and 3^{rd} peak (Fig. 4) the reliability is hardly sufficient for initializing a location of the second source within one frame. Nevertheless, the information of the second sound source position contained in the second maximum peak of the GCC function can be used: By localizing sound sources only with the first peak by means of the reliability criteria for the first peak described in Sect. 4, region of interests for the TDOAs of microphone pairs of all different sound sources are initialized. To continue the initialized sound track, the information of the second maximum peaks can be used to increase the number of estimates for averaging successive frames for a more precise TDOA estimation. This is implemented successfully in a real-time acoustic multi-source tracker with the microphone array of Fig. 1 increasing the number of correct estimates for averaging by 41.02% [44.68%] for the *Double-Talk* [*Extreme Double-Talk*] scenario and hence enhancing the accuracy of the simultaneously active sound sources.

7. CONCLUSIONS

In this paper the possibility of considering the second peak in the GCC function for TDOA estimation in a noisy and reverberant multi-source environment was studied. With two confidence criteria, the absolute value of the second maximum peak and the ratio between the 2^{nd} and the 3^{rd} peak in the GCC function, it is possible to evaluate the reliability of the current TDOA estimate of the second sound source. But the maximum confidence for highest values of the confidence criteria with about 50% is not sufficient to initialize with these estimates a second sound source reliably. The information about the second sound source can be used all the same in increasing the number of estimates for averaging for already initialized sound sources and hence rendering the TDOA estimates more precise.

To integrate the array into a humanoid robot, future work will be devoted to miniaturizing the array and to extend the system to multiple moving sources.

8. ACKNOWLEDGMENTS

This work is part of the Sonderforschungsbereich (SFB) No. 588 *"Humanoide Roboter - Lernende und kooperierende multimodale Roboter"* at the University of Karlsruhe Germany. The SFB is supported by the Deutsche Forschungsgemeinschaft (DFG).

9. REFERENCES

[1] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, August 1976.

[2] J.O. Smith and J.S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoustics, Speech, Signal Processing*, 1987.

[3] Y. Huang, J. Benesty, and G.W. Elko, "Passive acoustic source localization for video camera steering," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, June 2000, pp. 909–912.

[4] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein, *Microphone Arrays*, chapter Robust Localization in Reverberant Rooms, Springer, 2001.

[5] D.E. Sturim, M.S. Brandstein, and H.F. Silverman, "Tracking multiple talkers using microphone-array measurements," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, April 1997, vol. 1, pp. 371 –374.

[6] D. Bechler and K. Kroschel, "Confidence scoring of time difference of arrival estimation for speaker localization with microphone arrays," in *13. Konferenz Elektronische Sprachsignalverarbeitung ESSV*, Dresden, September 2002.

[7] D. Bechler and K. Kroschel, "Reliability measurement of time difference of arrival estimations for multiple sound source localization," in *17th Annual Meeting of the IAR*, Grenoble, November 2002.