INTERSPEECH 2009 BRIGHTON

Estimating the position and orientation of an acoustic source with a microphone array network

Alberto Yoshihiro Nakano, Seiichi Nakagawa, Kazumasa Yamamoto

Department of Information and Computer Sciences, Toyohashi University of Technology, Japan
{alberto, nakagawa, kyama}@slp.ics.tut.ac.jp

Abstract

We propose a method that finds the position and orientation of an acoustic source in an enclosed environment. For each of eight T-shaped arrays forming a microphone array network, the time delay of arrival (TDOA) of signals from microphone pairs, a source position candidate, and energy related features are estimated. These form the input for artificial neural networks (ANNs), the purpose of which is to provide indirectly a more precise position of the source and, additionally, to estimate the source's orientation using various combinations of the estimated parameters. The best combination of parameters (TDOAs and microphone positions) yields a 21.8% reduction in the mean average position error compared to baselines, and a correct orientation ratio higher than 99.0%. The position estimation baselines include two estimation methods: a TDOA-based method that finds the source position geometrically, and the SRP-PHAT that finds the most likely source position by spatial exploration.

Index Terms: microphone array network, position and orientation estimation, artificial neural network

1. Introduction

Microphone arrays [1] have received increasing attention in the past few years, especially for spatial filtering (beamforming)[2], and sound source localization for speech, audio, and acoustics processing. Acoustic localization is also an important task in many practical applications such as videoconferencing [3], hands-free communication systems [4], hearing aids [5], and human-machine interaction [6]. In a previous work [7], we tried to find the *best array*, defined as the one that yields the best position estimate, out of all the arrays in an array network, and additionally, to estimate the source orientation. For the task, we used an ANN with its input set composed of energy related features (power and correlation values) and a *distance* value defined as the Euclidean distance between the array and its estimated position candidate, and giving as its result the orientation and *best array* at the output. In this paper, instead of choosing the *best array* using *distance* information, we estimate a more accurate position using the position candidates set, the TDOAs set, and the microphone positions of all arrays in the network. This approach has the advantage of exploring more spatial information in 2D and 3D spaces compared to the one-dimensional information given by the *distance* information only.

For the estimation of TDOA and position candidates, a robust GCC-PHAT (generalized cross-correlation with phase transform) function [7] was employed. Position candidates were estimated by two different methods: a TDOA-based method [8] and the SRP-PHAT (steered response power with phase transform) [1]. In the former method, the optimal position is determined by geometric derivation and is highly dependent on the correct estimation of the time delays, whereas in the latter method, the optimal position is obtained by steering the microphone array to all potential source positions looking for the point in the space with the highest spatial likelihood. We note that using position candidates estimated by SRP-PHAT yielded better results than using those obtained by the TDOA-based method. However, the best results were obtained using a combination of TDOAs and microphone positions based on ANNs.

The outline of this paper is as follows. In Section 2, we briefly describe both the TDOA-based and the SRP-PHAT position estimation methods. In Section 3, we present the modification to our position and orientation estimation method from our previous work. In Sections 4 and 5, we discuss the experimental conditions and results, respectively, and we conclude in Section 6.

2. Background

In position estimation methods employed as baselines, a robust version of the GCC-PHAT function [7] was employed. In the TDOA-based method, the function was used to estimate the time delay of arrival of signals from microphone pairs, whereas in the SRP-PHAT method, it was used to create a spatial sound map of the test environment. Given below is a short description of each position localization method adopted as a baseline.

2.1. TDOA-based position estimation method

We used the position estimation method from [8] tailored for the T-shaped microphone array shown in Figure 1. In the initial step, using the robust GCC-PHAT function, a set of 3 TDOAs, $\{\tau_{12},\tau_{13},\tau_{14}\}$, is estimated for pairs $\{q_1, q_2\}$, $\{q_1, q_3\}$, $\{q_1, q_4\}$, taking q_1 as the reference, while in the second step, the source position is found by geometric derivation. The idea behind this method involves finding the intersection point of 3 hyperplanes in the space, with each one defined for each TDOA.

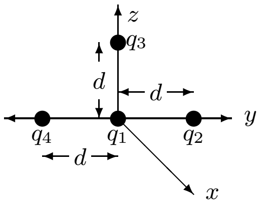

Figure 1: T-shaped microphone array composed by microphones $\{q_1, q_2, q_3, q_4\}$. d is the distance between adjacent microphones.

2.2. SRP-PHAT estimation method

The SRP-PHAT [1] is a robust position localization method that explores the space, searching for the region with the highest spatial likelihood obtained by a cumulative voting process involving cross-correlation functions of microphone pairs. In practice, the space is divided into small regions and the theoretical delays between these regions and microphone pairs are pre-computed and stored. Thus, each small region l, characterized by a point in the space $\boldsymbol{\alpha}_l = (x_l, y_l, z_l)$, is associated with a vector of time delays

$$\boldsymbol{\tau}(\boldsymbol{\alpha}_l) = [\tau_{12}(\boldsymbol{\alpha}_l), \tau_{13}(\boldsymbol{\alpha}_l), \ldots, \tau_{mn}(\boldsymbol{\alpha}_l)], \qquad (1)$$

where $m, n = 1, \ldots, Q$ for $m \neq n$. After the cross-correlation functions between microphone pairs have been calculated by a robust GCC-PHAT function, a search-and-sum procedure is performed. For each small region l, the cross-correlation values ($R(.)$) corresponding to the theoretical time delays $\boldsymbol{\tau}(\boldsymbol{\alpha}_l)$ are found and accumulated. Once all regions have been swept, an acoustic map is created in the space. Finally, it is assumed that the most likely source position $\hat{\boldsymbol{\alpha}}$ will be the region with the highest spatial likelihood. The SRP-PHAT method can be mathematically formulated as:

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}_l} \sum_{m \neq n} R(\tau(\boldsymbol{\alpha}_l)). \qquad (2)$$

In our experiments, we divide the space into small regions of 5.0 cm × 5.0 cm × 5.0 cm. In this work, we denote SRP-PHAT$_{Array}$ as the method in which one position estimate (position candidate) is obtained by each array, and SRP-PHAT$_{All}$ as the method in which one position estimate is obtained by the entire array network.

3. Proposed position and orientation estimation method

The proposed method is implemented using a two-stage ANN (Figure 2), where the outputs of the first and second stages are the orientation and position, respectively. Two different input sets were tested. In the first set, power, correlation, and position candidate estimates for every array were combined; whereas in the second set, power, correlation, TDOA, and microphone positions for every array were combined. The purpose of the ANN is to provide indirectly a more precise position of the source and, additionally, to estimate the source's orientation using different combinations of the estimated parameters.

The power of an array is defined as the highest power value for all microphones in the array, while the correlation of an array is defined as the highest correlation value for all microphone pairs in the array. For each array, we have a single value for power, a single value for correlation, 2 or 3 values representing the position candidates in 2D or 3D, respectively, 3 values for time delays corresponding to the same set used by the TDOA-based position estimation method, and 12 microphone values corresponding to 4 microphones coordinate values in 3D space. The sets of values for power, correlation, position candidates, TDOAs, and microphone positions for the entire network are denoted by "P" (8 values), "C" (8 values), $\{\hat{\mathrm{x}}, \hat{\mathrm{y}}\}/\{\hat{\mathrm{x}}, \hat{\mathrm{y}}, \hat{\mathrm{z}}\}$ in 2D/3D space (16/24 values), $\{\hat{\boldsymbol{\tau}}\}$ (24 values), and $[\mathrm{x,y,z}]_{pq_m}$ (96 values) for array $p = 1, \ldots, P$ and microphones q_1, q_2, q_3, q_4, respectively.

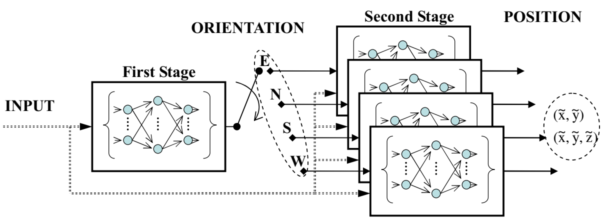
ORIENTATION
Second Stage
POSITION
First Stage
INPUT
E
N
S
W
$(\tilde{\mathrm{x}}, \tilde{\mathrm{y}})$
$(\tilde{\mathrm{x}}, \tilde{\mathrm{y}}, \tilde{\mathrm{z}})$

Figure 2: Two-stage artificial neural network topology used in this study. **INPUT** = $\{\mathrm{P} + \mathrm{C} + \{\hat{\mathrm{x}}, \hat{\mathrm{y}}\}$ / $\{\hat{\mathrm{x}}, \hat{\mathrm{y}}, \hat{\mathrm{z}}\}\}$; or **INPUT** = $\{\mathrm{P} + \mathrm{C} + \{\hat{\boldsymbol{\tau}}\} + [\mathrm{x,y,z}]_{pq_m}\}$; **1st STAGE OUTPUT** = {Orientation (E, N, S, W)}; **2nd STAGE OUTPUT** = {POSITION}.

4. Experimental setup

All experiments were conducted in a 5 m × 6.4 m × 2.65 m room. Eight T-shaped microphone arrays ($P = 8$) were used, with one array fixed to each wall (arrays A, B, C, and D) and four arrays fixed to the ceiling (arrays E, F, G, and H). Each array was mounted on a structure composed of acoustic absorber to reduce reflection effects near the microphones. The distance between pairs of microphones in each array was set to $20cm$. The room was divided into 50 areas, each one 50 cm by 50 cm, but only 29 areas, suitably distributed and covering the entire room, were considered in our analyses. The array positions and areas are depicted in Figure 3. A loudspeaker (ALR JORDAN) was set-up over a stand fixed 140 cm above the floor to simulate an acoustic source. The stand was centered in each area and 300 Japanese words were played, with the mean duration of an utterance being 0.6 s. In each studied area, the loudspeaker was turned to four different orientations shifted by 90°; east (E), north (N), south (S), and west (W) orientations were considered, resulting in 116 study cases (29 × 4). In the experiments, one position estimate was obtained per utterance.

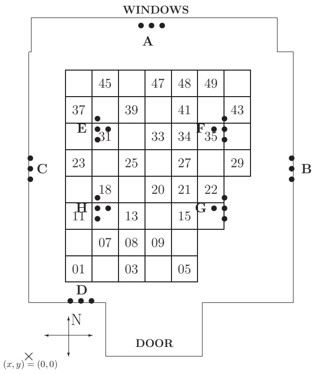

Figure 3: View of the room from above. The 29 studied areas are numbered, while the origin of the coordinate system, as well as the relative orientation are shown at the bottom left.

In the ANN analyses, fully connected feedforward ANNs were implemented using the Stuttgart Neural Network Simulator (SNNS) [1]. We considered results in both 2D and 3D space $(\tilde{\mathrm{x}}, \tilde{\mathrm{y}})$ / $(\tilde{\mathrm{x}}, \tilde{\mathrm{y}}, \tilde{\mathrm{z}})$. The ANN topology used in this study is illustrated in Figure 2. The first stage consists of a single ANN that estimates the source orientation, while the second stage consists of four individual ANNs, each one trained to consider one of the orientations E, N, S, or W. In the gating ANN at the first

[1] http://www.ra.cs.uni-tuebingen.de/SNNS/

stage, using a combination of power “P” values, correlation “C” values and $\{\hat{x}, \hat{y}\}$ / $\{\hat{x}, \hat{y}, \hat{z}\}$ values, the source orientation is estimated and used to select the corresponding ANN at the second stage whose output is $(\tilde{x}, \tilde{y})$ or $(\tilde{x}, \tilde{y}, \tilde{z})$, depending on whether the ANN was trained using 2D or 3D position candidates. Note that at the second stage, both the orientation information and the initial input set are required. The same observations are valid using “P”, “C”, $\{\hat{\boldsymbol{\tau}}\}$ and $[x, y, z]_{pq_m}$ as the input set, but considering only the 3D case. Table 1 presents the ANN configurations studied in this research.

For the ANN training/testing phase, recorded data from each of the 29 areas were divided into two sets with 80% of the data used in the training phase and 20% in the testing phase, and with no overlap between the training and testing data sets. Five different data sets were created by permuting all recorded data for cross-validation. Results are presented in terms of correct orientation ratio (%), mean average orientation error $(^o)$, and mean average position error (cm), where ratio is the relation of the total number of correct estimates by the ANN to the total number of input patterns, orientation error corresponds to the mismatch between the actual and estimated orientations, and the position error is the Euclidean distance between the estimated position and the actual source position. CLOSED and OPEN tests refer to results obtained by the trained ANN evaluated using the training and testing data sets, respectively. In the training phase, at the first stage, the correct orientation was used as the target value, while at the second stage the actual source position was used.

Table 1: Two-stage ANN configuration. The results were obtained considering the hidden unit numbers in bold. “4/2” or “4/3” in OUTPUT column denotes the number of outputs in the first and second stages, respectively.

NUMBER OF ANN UNITS		
INPUT	HIDDEN	OUTPUT
32 (P+C+$\{\hat{x}, \hat{y}\}$)	**80**	4/2
40 (P+C+$\{\hat{x}, \hat{y}, \hat{z}\}$)	**80**	4/3
120 ($[x, y, z]_{pq_m}$+$\{\hat{\boldsymbol{\tau}}\}$)	**240**	4/3
136 (P+C+$[x, y, z]_{pq_m}$+$\{\hat{\boldsymbol{\tau}}\}$)	**272**	4/3

5. Experimental results

Table 2 gives the comparative results for the position localization methods presented in Section 2. In this table, the TDOA-based and SRP-PHAT$_{Array}$ methods calculate position candidates for every array, and the mean average position error considers the case when the best estimate from all arrays is always selected, that is, using an *oracle* selection. SRP-PHAT$_{All}$ finds the position estimate using the whole microphone array network. These values are adopted as baselines for comparison with the performance of our proposed automatic system. In the *oracle* selection, SRP-PHAT$_{Array}$ is better than the TDOA-based method; however, SRP-PHAT$_{All}$ is the best estimation method because in using the entire network the exploration of the spatial properties of the sound field is not restricted to the characteristics of an individual microphone array.

Tables 3 and 4 present the results for the proposed automatic position and orientation estimation method using, respectively, position candidates from the TDOA-based and SRP-PHAT$_{Array}$ methods, and TDOAs and microphone positions. INPUTS, SPACE and TEST denote, respectively, the number of units in the ANN input layer, the dimensional space, that is, whether ANNs were trained using 2D or 3D position estimated data, and the test condition, that is, either an OPEN or CLOSED test.

Table 2: Mean average position error in centimeters for TDOA-based and SRP-PHAT$_{Array}$ estimation methods in *oracle* selection. In SRP-PHAT$_{All}$, all arrays are used in the estimation task. 2D (3D) represent two (three) dimensional space, respectively.

	SPACE	
Method	2D	3D
TDOA-based (*oracle*)	21.2	34.1
SRP-PHAT$_{Array}$ (*oracle*)	18.6	31.7
SRP-PHAT$_{All}$	**16.0**	**29.8**

In Table 3, we compare the performance of our ANN approach using position candidates from both position localization methods. Using TDOA-based position candidates in the ANN, we obtained a mean average position error of 31.8 cm, which is better than the TDOA-based (oracle) results of 34.1 cm in 3D space, and a correct orientation ratio around 90%. Using SRP-PHAT$_{Array}$ position candidates in the ANN, a mean average position error of 27.9 cm was obtained, which is better than all the baselines in 3D space, even the SRP-PHAT$_{All}$ with 29.8 cm. Moreover, a correct orientation ratio of more than 98% was obtained. The values marked by $(^*)$ are 2D estimates obtained directly from the 3D estimates but disregarding the z dimension, and these appear to be better estimates than those obtained by training the ANNs using the 2D data set. Comparing the results obtained using position candidates from the TDOA-based and SRP-PHAT$_{Array}$ methods in the ANNs, it is evident that SRP-PHAT$_{Array}$ yields better results, because the TDOA-based method is highly dependent on the correct estimation of time delays, and it is quite difficult to estimate these precisely in a real environment.

Table 4 presents the results using the energy related features, TDOAs, and microphone positions in the ANN. In the columns marked “CLOSED” (`CLOSED POSITIONS TEST`), the training and testing data sets were obtained as in Table 3. In the columns marked “OPEN” (`OPEN POSITIONS TEST`) the simulation conditions were modified to separate the training and testing data sets by area. Six areas were randomly chosen inside the room, and avoiding border areas. (Areas 1, 3, 5, 11, 22, 23, 29, 37, 43, 45, 47, and 49 are border areas in Figure 3. Although area 48 is also a border area, it is surrounded by other areas and can thus be disregarded.) The data from these areas were used as the testing data set, while the data from the other 23 areas formed the training data set. There was no overlap between the training and testing data sets, and five different data permutations were simulated for cross-validation. Based on this separation, approximately 80% and 20% of the data were used as training and testing data sets, respectively.

Comparing Tables 4 and 3, it appears that the TDOA estimates and microphone positions are more suitable as input for the ANNs for estimating the orientation and position than the position candidates. In `CLOSED POSITIONS TEST`, an almost perfect orientation estimation of 99.5% in the correct orientation ratio, and position estimation of 23.3 cm in 3D in the OPEN test case were obtained, which reflects a reduction of 21.8% compared to the baseline SRP-PHAT$_{All}$. However, SRP-PHAT$_{All}$ (16.0 cm) is still better in 2D space compared to this method (20.5 cm). In `OPEN POSITIONS TEST`, the

Table 3: Results obtained by the input set $\{\mathrm{P+C+}\{\hat{\mathrm{x}},\hat{\mathrm{y}}\}/\{\hat{\mathrm{x}},\hat{\mathrm{y}},\hat{\mathrm{z}}\}\}$. Position candidates $\{\hat{\mathrm{x}},\hat{\mathrm{y}}\}/\{\hat{\mathrm{x}},\hat{\mathrm{y}},\hat{\mathrm{z}}\}$ in 2D/3D were estimated by **TDOA-based** (TDOA) and **SRP-PHAT**$_{Array}$ (SRP) methods. CLOSED/OPEN means that the training/testing data set was used to evaluated the ANN. The mean average position error value was calculated in the considered dimensional space, and the value (*) was the mean average position error in 2D directly calculated from 3D estimates.

			RESULTS					
			Corr. orient.		Mean avg.		Mean avg.	
			ratio (%)		orient. error (°)		pos. error (cm)	
INPUTS	SPACE	TEST	TDOA	SRP	TDOA	SRP	TDOA	SRP
$32(\mathrm{P+C+}\{\hat{\mathrm{x}},\hat{\mathrm{y}}\})$	2D	CLOSED	87.4	97.6	15.7	2.9	29.9 (27.9*)	23.1 (23.1*)
$32(\mathrm{P+C+}\{\hat{\mathrm{x}},\hat{\mathrm{y}}\})$	2D	OPEN	83.9	94.7	19.7	6.5	31.1 (29.1*)	23.5 (23.5*)
$40(\mathrm{P+C+}\{\hat{\mathrm{x}},\hat{\mathrm{y}},\hat{\mathrm{z}}\})$	3D	CLOSED	93.4	99.7	8.2	0.4	32.9	27.6
$40(\mathrm{P+C+}\{\hat{\mathrm{x}},\hat{\mathrm{y}},\hat{\mathrm{z}}\})$	3D	OPEN	90.3	**98.3**	12.0	2.2	33.9	**27.9**

Table 4: Results obtained by the input set $\{\mathrm{P}+\mathrm{C}+\{\hat{\boldsymbol{\tau}}\}+[\mathrm{x,y,z}]_{pq_m}\}$. TDOAs and microphone positions of every array were employed. *CLOSED*/*OPEN* states for closed/open position conditions. CLOSED/OPEN means that the training/testing data set was used to evaluated the ANN. The value (*) was the mean average position error in 2D directly calculated from 3D estimate.

		RESULTS					
		Corr. orient.		Mean avg.		Mean avg.	
		ratio (%)		orient. error (°)		pos. error (cm)	
INPUTS	POSITION TEST	*CLOSED*	*OPEN*	*CLOSED*	*OPEN*	*CLOSED*	*OPEN*
$120([\mathrm{x,y,z}]_{pq_m}+\{\hat{\boldsymbol{\tau}}\})$	CLOSED	99.9	96.8	0.1	4.2	24.4 (20.9*)	24.9 (21.6*)
$120([\mathrm{x,y,z}]_{pq_m}+\{\hat{\boldsymbol{\tau}}\})$	OPEN	99.4	**87.0**	0.7	17.0	24.6 (21.0*)	**25.3** (22.5*)
$136(\mathrm{P+C+}[\mathrm{x,y,z}]_{pq_m}+\{\hat{\boldsymbol{\tau}}\})$	CLOSED	99.9	96.3	0.1	4.5	23.2 (20.3*)	25.7 (22.1*)
$136(\mathrm{P+C+}[\mathrm{x,y,z}]_{pq_m}+\{\hat{\boldsymbol{\tau}}\})$	OPEN	**99.5**	84.2	0.6	18.5	**23.3** (20.5*)	28.2 (24.8*)

objective was to estimate jointly the position and orientation in untrained areas. The results show a correct orientation ratio range from 84% to 87% and a mean average position error of 25.3 cm, which is still better than the baseline SRP-PHAT$_{All}$ in 3D space.

Just for comparison, in our previous work [7] using the *distance* information, we obtained values for correct orientation ratio and mean average position error of 77.8% (76.0%) and 32.6 cm (45.8 cm), respectively, in 2D (3D) space for the TDOA-based method, and of 79.3% (76.8%) and 28.5 cm (43.4 cm), respectively, in 2D (3D) space for the SRP-PHAT$_{Array}$ method under OPEN test conditions. The improvement in the results from this work can be explained by the fact that 2D or 3D coordinates have more spatial information than a one-dimensional value defined by *distance* information that leads to unreliable estimation values. In other words, using *distance* values derived from position estimation candidates there is a loss of useful information.

6. Conclusions

In this work, we expanded our research on position and orientation estimation of an acoustic source. Exploring spatial information provided by position candidates derived from each array, TDOAs between microphone pairs in each array, and microphone positions in the array network, we achieved a significant improvement in the position and orientation estimations. The proposed method could be applicable, for instance, in position dependent speech recognition tasks [9].

7. Acknowledgements

We would like to thank the Global COE program "Frontiers of Intelligent Sensing" and Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT) for supporting our research.

8. References

[1] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. New York: Springer, 2001.

[2] J. G. Ryan and R. A. Goubran, "Optimum near-field performance of microphone arrays subject to a far-field beampattern constraint," *Journal of Acoustical Society of America*, vol. 108, pp. 2248–2255, November 2000.

[3] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system invideoconferencing," *Proceedings of ICASSP*, vol. I, pp. 187–190, 1997.

[4] S. Fischer and K. U. Simmer, "An adaptive microphone array for hands-free communication," *in Proc. 4th International Workshop on Acoustic Echo and Noise Control, IWAENC-95*, pp. 44–47, 1995.

[5] M. R. Bai and C. Lin, "Microphone array signal processing with application in three-dimensional spatial hearing," *J. of Acoustical Society of America*, vol. 117, pp. 2112–2121, April 2005.

[6] K. Nakadai, H. Nakajima, M. Murase, S. Kaijiri, K. Yamada, T. Nakamura, Y. Hasegawa, H. G. Okuno, and H. Tsujino, "Robust tracking of multiple sound sources by spatial integration of room and robot microphone arrays," *Proceedings of ICASSSP*, pp. IV 929–932, 2006.

[7] A. Y. Nakano, K. Yamamoto, and S. Nakagawa, "Directional acoustic source's position and orientation estimation approach by a microphone array network," *in IEEE Proc. of the 13th DSP Workshop & 5th SPE Workshop*, pp. 606–611, January 2009.

[8] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distance speaker recognition based on position-dependent CMN by combing speaker-specific GMM with speaker-adapted HMM," *Speech Communications*, vol. 49, pp. 501–513, 2007.

[9] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speech recognition by combining position-dependent CMN with conventional CMN," *Proceedings of ICASSP*, vol. 4, pp. 817–820, 2007.