Automatic estimation of position and orientation of an acoustic source by a microphone array network

Alberto Yoshihiro Nakano,[a] Seiichi Nakagawa, and Kazumasa Yamamoto
Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi 441-8580, Japan

(Received 24 March 2009; revised 29 September 2009; accepted 2 October 2009)

A method which automatically provides the position and orientation of a directional acoustic source in an enclosed environment is proposed. In this method, different combinations of the estimated parameters from the received signals and the microphone positions of each array are used as inputs to the artificial neural network (ANN). The estimated parameters are composed of time delay estimates (TDEs), source position estimates, distance estimates, and energy features. The outputs of the ANN are the source orientation (one out of four possible orientations shifted by 90° and either the best array which is defined as the nearest to the source) or the source position in two dimensional/three dimensional (2D/3D) space. This paper studies the position and orientation estimation performances of the ANN for different input/output combinations (and different numbers of hidden units). The best combination of parameters (TDEs and microphone positions) yields 21.8% reduction in the average position error compared to the following baselines and a correct orientation ratio greater than 99%. Position localization baselines consist of a time delay of arrival based method with an average position error of 34.1 cm and the steered response power with phase transform method with an average position error of 29.8 cm in 3D space.

© *2009 Acoustical Society of America.* [DOI: 10.1121/1.3257548]

PACS number(s): 43.60.Jn, 43.60.Np, 43.66.Qp [EJS] Pages: 3084–3094

I. INTRODUCTION

Microphone arrays[1,2] have received increasing attention in the past few years, especially in spatial filtering (beamforming)[3] and sound source localization[4–6] for speech, audio, and acoustic processing. Microphone array techniques depend on many factors, including placement, geometrical configuration, number of microphones, as well as the conditions and the number of active acoustic sources in the environment under investigation. This is the main reason why a distributed microphone array network, comprised of eight T-shaped microphone arrays, is used in this study to estimate the position and orientation of a directional acoustic source in an actual enclosed environment.

Acoustic localization is an important task in many practical applications such as videoconferencing,[7] hands-free communication system,[8] hearing aids,[9] and human-machine interaction.[10] Different source localization methods were proposed in the literature[1,6,11–13] and they can be separated into two groups; indirect and direct estimation methods. The difference consists of whether a set of time delay estimates (TDEs) of microphone pair signals is estimated or not and used in the localization method.

Average magnitude difference function (AMDF),[14] adaptive eigenvalue decomposition (AED),[11] information theory (IT),[15] and generalized cross-correlation[16] (GCC) methods are some examples of time delay estimation techniques. In the AMDF approach, an average magnitude difference function is calculated from the microphone signals, and the time delay corresponds to the time lag which minimizes this function. In this approach, when the signals are time aligned, the average magnitude difference between them becomes minimum. The AED approach focuses directly on the estimation of the impulse responses between the acoustic source and a pair of microphones. It is assumed that the relative delay between signals corresponds to the time lag between direct path components of estimated impulse responses. In the IT approach, it is shown that the time delay that maximizes the mutual information between pairs of signals is equivalent to the time delay that maximizes the cross-correlation between these signals. The GCC approach has been the traditional method for time delay estimation, where the optimum time delay corresponds to the time lag which maximizes the cross-correlation function obtained for a pair of microphone signals. A variant of the GCC approach, the GCC with phase transform (GCC-PHAT), is employed in this research. In the GCC-PHAT method, each component of the cross-power spectrum of microphone pair signals is equally emphasized for generating a prominent peak in the cross-correlation function. But, due to the noise and reverberation conditions in the test environment, a more robust GCC-PHAT is derived in Sec. II C and used for both the indirect and direct position estimation methods employed in this research.

In indirect localization methods, a set of time delays is initially estimated from different microphone pair signals, and the optimal source position is found by geometrical derivation, that is, by solving a set of equations that compute the intersection point of hyperplanes in the space, with each hyperplane defined for each TDE. Examples of these methods can be found in Ref. 12, originally derived for global posi-

[a] Author to whom correspondence should be addressed. Electronic mail: alberto@slp.ics.tut.ac.jp

tioning system but useful for acoustic localization when the array geometry is unknown, and in Ref. 6 when the array configuration is known *a priori*, which permits simplifications in the formulation of the position estimation method. Indirect methods are highly dependent on the correct delay estimation, and either inaccurate estimates or small variations in the estimates to the optimum values imply an unreliable position estimate that can be outside of the studied space. To avoid mismatched microphones due to production tolerances, aging effects, and unbalanced microphone array geometry, array and microphone compensation methods were proposed[2,17,18] in order to reduce the estimate variance. In this work, it is assumed that all microphones and arrays are well balanced, disregarding mismatch errors.

Direct localization methods are space exploration techniques that search the studied space for the point with highest spatial likelihood. Steered response power phase transform[1] (SRP-PHAT), based on the maximization of power obtained by steering the microphone array to all potential source positions, and global coherence field (GCF),[19] similar to SRP-PHAT approach, are examples of these methods. Direct methods do not depend on the array geometry but require more computation compared to indirect methods.

The source orientation[19–23] also plays an important role in acoustic localization because a directional source does not radiate uniformly in all directions, and the quality of signals recorded by distant microphones is affected not only by environmental noise and reverberation but also by the speaker's relative orientation.[21] Sachar *et al.*[20] proposed the *energy method*, where differences in the source radiation pattern can be detected and used to predict the source orientation. Signals obtained by a huge microphone array (HMA), composed of 448 microphones, are processed (to compensate the inverse-square-law attenuation of the source signal, to reduce the background noise and masking effect of reverberation, and to enhance directional components of the signal) and used to estimate the source radiation pattern. An analysis of this pattern indicates the most likely source orientation having higher gain in the estimated radiation pattern compared to the other orientations. This method presented 60% correct orientation ratio at ±5° tolerance level or nearly 100% correct performance at ±10° tolerance. Abad *et al.*[21] presented the SRP-PHAT orientation estimation and the high/low band ratio (HLBR) orientation estimation methods. In the SRP-PHAT orientation approach, the position and orientation are estimated together, and in addition to the theoretical time delays computed from each spatial position to each microphone pair, weights representing the influence of each cross-correlation in terms of the relative orientation are computed from a normalized source directivity pattern. For this method, an average error about 44° was obtained.[21] The ability of the system to correctly classify the source orientation within eight classes shifted by 45° was about 37% and admitting a classification error of ±1 adjacent class a value of 73% was obtained. HLBR is defined as the ratio between energies of high (3500–4500 Hz) and low (200–400 Hz) frequency bands of the radiation pattern. It is shown that HLBR maintains the same characteristics of the radiation pattern and is more robust to array calibration mismatches and propagation losses. The angle which maximizes the correlation between a mathematical HLBR model and an estimated HLBR model corresponds to the orientation of the source. Using the same standard metrics of the SRP-PHAT orientation method, i.e., the average error, the ability of the system to correctly classify the source orientation within eight classes separated by 45°, and assuming correct classification error of ±1 adjacent class, 51°, 30%, and 68% were obtained, respectively. Brutti *et al.*[19,23] extended the GCF position localization method to consider the source orientation. The new method was named the oriented global coherence field. In this method, the exploration space is composed of weights corresponding to the position of the source and weights corresponding to the orientation of the source. The point in the space with highest sum of weights will show the estimated source position and the estimated orientation. It was reported that more than 99% of estimates were within an error of 45°.

In the studies on position and orientation estimation methods surveyed above, it is assumed that either the position is known *a priori* for orientation estimation or that the sound source is direct to the array so that the orientation is known *a priori* for position estimation. A few studies, as can be found in Refs. 21 and 23, estimate the position and the orientation together.

In this work, artificial neural networks (ANNs) are used in the position and orientation estimation. ANN (Ref. 24) is a massively parallel distributed structure with input-output mapping, it has the ability to learn from and adapt to certain conditions, and it can invoke assumptions about the underlying physical phenomena responsible for the generation of the input data. Employing ANNs, it is expected that combining estimated parameters of each array of the network could be sufficient to predict the source orientation and could give a more accurate source position than that estimated by an individual array. Here in this work, we define representation level as the compactness or aggregation of information of a given process, in our case the estimation of the position and the orientation, such that the higher the representation is, the smaller the number of parameters necessary to describe the same information will be. Figure 1 shows the representation level of the estimated parameters used at the ANN input. The low representation is given by TDEs and microphone positions, the intermediate representation is given by source position estimates, and the high representation parameter is given by distance estimates derived from source position estimates. There is a reduction in the number of estimated parameters from the lower to the higher representation level. In this work, for each array TDEs are represented by three values, microphone positions are represented by 12 values, a source position estimate is represented by three values, and a distance estimate is represented by one value. In Sec. IV E 2, we show that the lower the representation level is, the better are the ANN results.

The outline of this paper is organized as follows. In Sec. II classical and robust GCC-PHAT methods for TDEs are presented as well as two baseline position estimation methods; the time delay of arrival (TDOA)-based and the SRP-PHAT. In Sec. III, the proposed method that jointly estimates

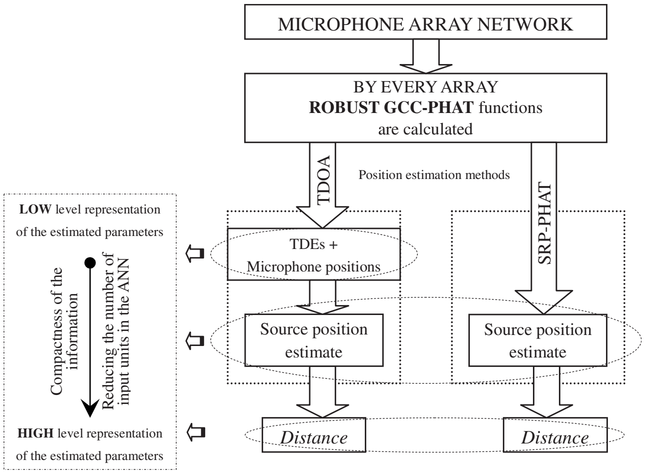
MICROPHONE ARRAY NETWORK
BY EVERY ARRAY
ROBUST GCC-PHAT functions
are calculated
TDOA
Position estimation methods
SRP-PHAT
LOW level representation
of the estimated parameters
Compactness of the information
Reducing the number of input units in the ANN
HIGH level representation
of the estimated parameters
TDEs +
Microphone positions
Source position estimate
Source position estimate
Distance
Distance

FIG. 1. Description of the estimation of different parameters (TDEs, source position estimate, and distance estimate) by every array of the microphone array network used in the proposed position and orientation estimation method.

the position and orientation of a directional acoustic source using ANNs is presented. In Sec. IV, experimental setup and results are presented, and in Secs. V and VI, discussions and conclusions are presented.

II. BACKGROUND

A. Signal model

Consider P identical arrays, each one with Q microphones, where each microphone is defined as m, for $m = 1, \ldots, Q$. Given a signal source $s(t)$, the signal at each microphone can be represented as

$$\mathsf{x}_{p_m}(t) = h_{p_m}(t) * s(t) + n(t), \tag{1}$$

where $p \in \{1, \ldots, P\}$, "$*$" denotes convolution, $h_{p_m}(t)$ is the reverberation impulse response that describes the propagation path between the source $s(t)$ and the mth microphone of the pth array, and $n(t)$ is the additive background noise.

B. GCC-PHAT method and time delay estimation

The GCC-PHAT function calculated from two signals $\mathsf{x}_{p_m}(t)$ and $\mathsf{x}_{p_n}(t)$, $m \neq n$, of the pth array is defined as

$$R(\tau_{p_{mn}}) = \int_{-\infty}^{+\infty} \frac{\mathsf{X}_{p_m}(f)\mathsf{X}^*_{p_n}(f)}{|\mathsf{X}_{p_m}(f)\mathsf{X}^*_{p_n}(f)|} e^{-j2\pi f \tau_{p_{mn}}} df, \tag{2}$$

where $R(\tau_{p_{mn}})$ is a function of the time delay $\tau_{p_{mn}}$, and $\mathsf{X}_{p_m}(f)$ and $\mathsf{X}_{p_n}(f)$ are the spectral representations of the signals $\mathsf{x}_{p_m}(t)$ and $\mathsf{x}_{p_n}(t)$, respectively. In this method, a phase dependent equalized cross-power spectral function becomes a time difference dependent cross-correlation function by Fourier transform. Therefore, the TDE $\hat{\tau}_{p_{mn}}$ between pair of signals corresponds, ideally, to the time difference that maximizes $R(\tau_{p_{mn}})$ as

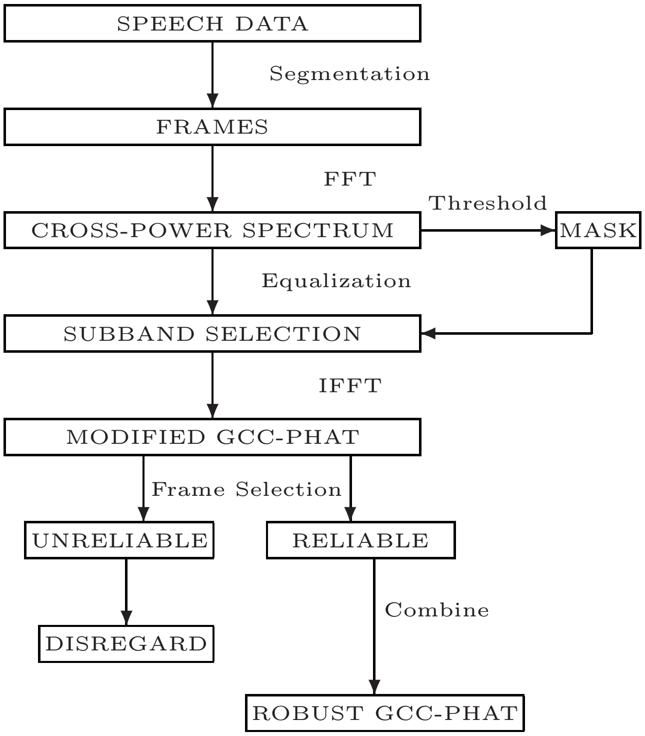

FIG. 2. Diagram block which illustrates the improvement of GCC-PHAT method.

$$\hat{\tau}_{p_{mn}} = \max_{\tau_{p_{mn}}}\{R(\tau_{p_{mn}})\}. \tag{3}$$

In the equalization procedure in Eq. (2), all frequency components are equally weighted generating a prominent peak in time domain. However, components of the spectrum with weak signal-to-noise ratio (SNR) are also accentuated, reducing the efficiency of this method under high noise conditions.[1] In Sec. II C, this fact is explored to generate a robust GCC-PHAT function. The TDEs are used in the localization algorithm presented in this work.

C. Improving GCC-PHAT function

Steps to improve the GCC-PHAT function are described in the block diagram of Fig. 2. Initially, signals $\mathsf{x}_{p_m}(t)$ and $\mathsf{x}_{p_n}(t)$ are segmented into L frames of N samples. Each pair of frames is converted to its frequency domain representation, and the cross-power spectral function is calculated. In the next stage, a subband selection method, where only high SNR spectrum components are selected in the cross-power spectral function, is performed. In this method a threshold γ is defined and a binary mask is created (see Fig. 3) as follows:

$$\text{mask} = \begin{cases} 1, & b > \gamma \\ 0 & \text{otherwise}, \end{cases} \tag{4}$$

where b is the magnitude of the cross-power spectrum component value. γ is defined as

$$N\,.\,\text{level} \leqslant \gamma < N\,.\,\text{level} + G \times \text{peak}, \tag{5}$$

where N.level, G, and peak are the average noise level, a gain factor $0 < G < 1$, and the highest cross-power peak

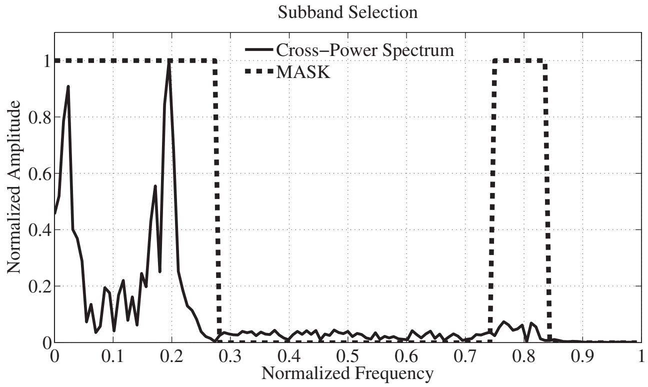

FIG. 3. Subband selection method. Binary mask (dashed line) selects the cross-power spectrum subbands (continuous line) with high SNR.

value in the frame, respectively. A value of $G=0.05$, used in the experiments, was determined experimentally by analyzing our recorded files. Then, the generated mask is applied to the equalized cross-power spectral function to reduce parts of undesirable contributions of noise and reverberation, and the inverse Fourier transform converts the masked cross-power spectral function to a modified cross-correlation function. At this point, the combination of the estimated cross-correlation function of different frames could generate a robust cross-correlation function; however, not all frames will contribute positively in this procedure due to the low SNR and strong and long reverberation effects which vary from frame to frame. A way to select reliable frames, i.e., frames that improve the GCC-PHAT function, is needed. In the procedure adopted in this work, the TDEs are calculated by frames using aforementioned modified cross-correlation functions for all microphone pairs of an array. If all TDEs have zero values, then it is assumed that the TDEs were generated by *unreliable frames*. If the time delays have zero values over all microphone pairs, then the propagated wave arrives in phase at the array. This situation is only possible when the source is located far away, although, in reality the source is at a relative near distance from the considering microphone array. Unreliable frames were due to the contribution of the undesirable correlated noise at the array. Finally, disregarding *unreliable frames* and combining only *reliable frames*, a robust GCC-PHAT function is created

$$R'(\tau_{p_{mn}})=\sum_{k=1}^{K} R(\tau_{p_{mn}},k), \qquad (6)$$

where k is the index of the set of reliable frames, $k\in[1,\ldots,K]$, so that K, $K<L$, is the total number of reliable frames, $R(\tau_{p_{mn}},k)$ is the modified GCC-PHAT function of the kth reliable frame, and $R'(\tau_{p_{mn}})$ is the robust GCC-PHAT function. The TDE $\hat{\tau}_{p_{mn}}$ can be obtained as in Eq. (3) by maximizing $R'(\tau_{p_{mn}})$. An interpolation procedure[25] can be applied to the GCC-PHAT function to increase the resolution time lag, thereby improving the TDE accuracy. The search of the highest peak in GCC-PHAT function can be restricted by the array geometry and the studied space[25,26] to look only for physically possible delays, that is,

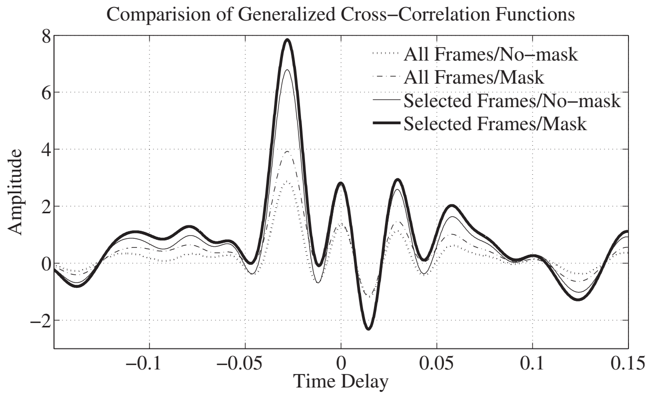

FIG. 4. GCC-PHAT function estimation with/without masking and/or frame selection. The dotted, dashed-dotted, thin solid, and thick solid lines represent the combination of all frames without masking, all frames with masking, only reliable frames without masking, and only reliable frames with masking of a segment of speech in the GCC-PHAT function estimation, respectively.

$$\tau_{p_{mn}}^{\min} \leqslant \hat{\tau}_{p_{mn}} \leqslant \tau_{p_{mn}}^{\max}, \qquad (7)$$

where $\tau_{p_{mn}}^{\min}$ and $\tau_{p_{mn}}^{\max}$ are the minimum and maximum possible values of $\tau_{p_{mn}}$. Figure 4 illustrates the evolution of the GCC-PHAT function considering the masking effect and/or reliable frame selection using the same segment of speech. It can be noted that with masking and frame selection a more prominent peak can be obtained.

D. TDOA-based and SRP-PHAT localization methods

TDOA based localization methods are performed in two steps: first, the TDEs of microphone pair signals are determined; and second, the source position estimate is found by the intersection point of hyperplanes in the space, with each one defined for each TDE. The intersection point is found by solving a set of defined equations. We used the position localization method from Wang *et al.*[6] tailored for the T-shaped microphone array shown in Fig. 5. The set of equations can be found in more detail in the cited reference. Thus, in the initial step, using the robust GCC-PHAT function described in Sec. II C, a set of three TDEs, $\{\tau_{12},\tau_{13},\tau_{14}\}$, is estimated for microphone pairs {1,2}, {1,3}, and {1,4}, taking microphone 1 as the reference, and in the second step, the source position estimate is found.

The SRP-PHAT 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

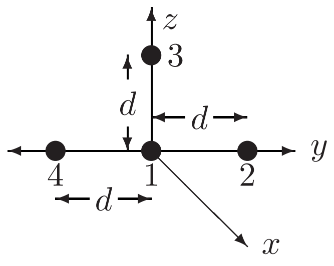

FIG. 5. T-shaped microphone array composed of microphones {1,2,3,4}. *d* is the distance between adjacent microphones.

pair signals. While in Ref. 1 this method is described in detail, here, we give only the basic concepts. In the SRP-PHAT method, the space is divided into small regions, and the theoretical delays between these regions and microphone pairs are pre-computed and stored. In our experiments, we divide the space into small regions of $5.0\times5.0\times5.0$ cm^3. 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

$$\boldsymbol{\tau}(\boldsymbol{\alpha}_l)=[\tau_{12}(\boldsymbol{\alpha}_l),\tau_{13}(\boldsymbol{\alpha}_l),\ldots,\tau_{1Q}(\boldsymbol{\alpha}_l)], \tag{8}$$

where $m,n=1,\ldots,Q$ for $m\neq n$. After the cross-correlation functions have been calculated by the robust GCC-PHAT function described in Sec. II C, a search-and-sum procedure is performed. For each small region l, the cross-correlation values 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

$$\hat{\boldsymbol{\alpha}}=\arg\max_{\boldsymbol{\alpha}_l}\sum_{m\neq n}R'(\tau(\boldsymbol{\alpha}_l)). \tag{9}$$

In this work, we denote SRP-PHAT$_{\text{array}}$ as the method in which one source position estimate is obtained by a single array, and SRP-PHAT$_{\text{all}}$ as the method in which one source position estimate is obtained by the entire array network.

III. PROPOSED SOURCE'S POSITION AND ORIENTATION ESTIMATION METHOD

The ANN is a structure composed of units (input, hidden, and output) connected by weighted links between them. If the unit has only output connections, then it is an input unit. If the unit has only input connections, then it is an output unit. If it has both types of connection, then it is a hidden unit. Each unit computes the sum of weighted outputs of prior units with connection leading to this unit, v, and applies to the activation function $f_{\text{act}}(\cdot)$, which limits the output amplitude. The logistic function

$$f_{\text{act}}(v)=\frac{1}{1+e^{-v}} \tag{10}$$

was assumed as the activation function whose output is in the interval [0, 1]. The output function takes the activation function output to generate the output of the unit. Here, the output function is the identity function. In this work, a gating two-stage ANN is studied in the estimation of the position and orientation.

Two-stage fully connected feedforward ANNs, illustrated in Fig. 6, are used in the joint estimation of the position and orientation of an acoustic source. Different combinations of the estimated parameters from the received signals and microphone positions of each array of the array network are used as input of different ANN configurations with different numbers of hidden units, as defined in Table III. We tested different numbers of hidden layers for each input/output configuration. By increasing the number of hidden units we could obtain better results but at the cost of increasing the adaptation time. Then, we opted to define the number of hidden units at most twice the number of the input units. The estimated parameters, defined in detail in Sec. III A, consist of energy features, TDEs, source position estimates, and distance estimates. We aimed to obtain a more accurate source position estimate when source position estimates from the TDOA-based or SRP-PHAT localization method are used as the source position starting values for the ANN algorithm. In the proposed method, the use of a HMA (Ref. 20) configuration, formed by hundreds of microphones, and analyses of theoretical radiation pattern models are avoided at the expense of a training phase in the ANN. A final consideration must be pointed out, here two different situations are investigated: the *best array selection* with orientation estimation and *position estimation* with orientation estimation.

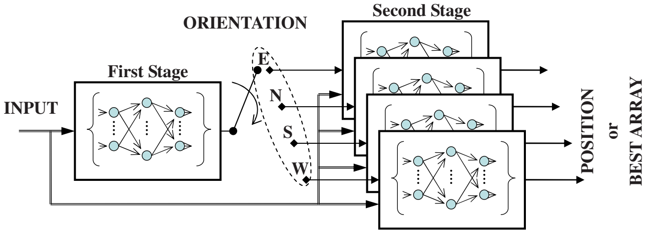

FIG. 6. (Color online) Two-stage artificial neural network studied topology. The configurations ANN$_1$, ANN$_2$, and ANN$_3$ are defined in Table III. The **INPUT**(ANN$_1$)={P+C+D} for the ANN$_1$ configuration comprises the set of energy features and distance estimates. The **INPUT**(ANN$_2$)={P+C+$\{\hat{x},\hat{y}\}/\{\hat{x},\hat{y},\hat{z}\}$} for the ANN$_2$ configuration comprises the set of energy features and source position estimates. The **INPUT**(ANN$_3$)={P+C+$\{\hat{\boldsymbol{\tau}}\}$+[x,y,z]} for the ANN$_3$ configuration comprises the set of energy features, TDEs, and microphone positions. The **first STAGE OUT.**={Orientation(E,N,S,W)} gives the orientation estimation at the first stage output and the **second STAGE OUT.**={(ANN$_1$)→BEST ARRAY; (ANN$_2$,ANN$_3$)→POSITION} gives either the best array or the source position estimate at the second stage output.

A. ANN input/output definitions

Energy features correspond to power and correlation estimates calculated by the recorded data. The maximum power at the array is defined as the highest power value over all signals of an array, that is,

$$P_p=\max_m\{P_{p_m}\}, \tag{11}$$

where P_{p_m} is the power calculated for signal $\mathsf{x}_{p_m}(t)$ of the pth array.

The maximum signal correlation across the array is defined as the highest correlation value over all signals of an array, that is,

$$C_p=\max_{m\neq n}\{C_{p_m,p_n}\}, \tag{12}$$

where C_{p_m,p_n} is the correlation calculated between signals $\mathsf{x}_{p_m}(t)$ and $\mathsf{x}_{p_n}(t)$ of the pth array.

Distance is defined by the Euclidean distance between the source position estimate and the corresponding array position which generated that estimate and is given by

$$D_p = \sqrt{(\hat{x}-x_p)^2 + (\hat{y}-y_p)^2 + (\hat{z}-z_p)^2}, \qquad (13)$$

where $(\hat{x},\hat{y},\hat{z})_p$ and (x_p,y_p,z_p) are the source position estimate by the pth array and the true position of the pth array, respectively. The position error, used to define the *best array*, is defined by the Euclidean distance between the source position estimate and the real position of the source given by

$$\mathrm{Error}_p = \sqrt{(\hat{x}-x_0)^2 + (\hat{y}-y_0)^2 + (\hat{z}-z_0)^2}, \qquad (14)$$

where (x_0,y_0,z_0) is the real position of the source. In order to simplify the notation at tables and figures, P=$\{P_1,\ldots,P_P\}$, C=$\{C_1,\ldots,C_P\}$, D=$\{D_1,\ldots,D_P\}$, $\{\hat{\mathrm{x}},\hat{\mathrm{y}},\hat{\mathrm{z}}\}$=$\{(\hat{x},\hat{y},\hat{z})_1,\ldots,(\hat{x},\hat{y},\hat{z})_P\}$, $\{\hat{\mathrm{x}},\hat{\mathrm{y}}\}$=$\{(\hat{x},\hat{y})_1,\ldots,(\hat{x},\hat{y})_P\}$, [x,y,z] and $\{\hat{\boldsymbol{\tau}}\}$=$\{\hat{\boldsymbol{\tau}}_1,\ldots,\hat{\boldsymbol{\tau}}_P\}$ are adopted for the set of power estimates, correlation estimates, distance estimates, source position estimates in three dimensional (3D), source position estimates in two dimensional (2D), microphone positions, and set of TDEs for arrays $p=1,\ldots,8$ and microphones $m=1,2,3,4$, respectively.

B. Orientation and *best array* selection using distance estimates

Best array is the array which yields the most reliable source position estimate, it yields the smallest position error, and it is the closest to the true source position. In an attempt to automatically select the best array, the distance estimate D_p of the pth array was used. The set of distance estimates D (a total of eight values) with the set of energy features P (power) and C (correlation) was used in the input of the ANN whose output corresponded to the orientation and the best array. The tested ANN corresponds to the ANN$_1$ configuration defined in Sec. IV C. This experiment was evaluated to verify if the distance estimates could be used to select the best array.

C. Orientation and position estimation using source position estimates or TDE sets

At the input of the ANN, the set of energy features P and C and the set of source position estimates $\{\hat{\mathrm{x}},\hat{\mathrm{y}}\}$ or $\{\hat{\mathrm{x}},\hat{\mathrm{y}},\hat{\mathrm{z}}\}$ are directly used in a mapping process to find a more accurate source position estimate and estimate the orientation using the entire microphone array network information. This approach has the advantage of exploring the spatial information given by the set of source position estimates compared to one-dimensional information generated by the distance estimate set. The tested ANN corresponds to the ANN$_2$ configuration defined in Sec. IV C. When using TDE set $\{\hat{\boldsymbol{\tau}}\}$, an additional information given by the microphone position set [x,y,z] is used in the input. In this case, the tested ANN corresponds to the ANN$_3$ configuration defined in Sec. IV C.

IV. EXPERIMENTS AND RESULTS

A. Automatic estimation: Setup and conditions

All experiments were conducted in a 5 m wide × 6.4 m long × 2.65 m high room containing eight T-shaped microphone arrays (see Fig. 7), 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 with an acoustic absorber to reduce reflection effects near the microphones. The distance between adjacent microphones at each array was set to 20 cm. The position of the center microphone in each array is given in centimeters as follows: A(236.5, 619.0, 206.0); B(497.0, 354.5, 200.0); C(3.0, 354.5, 200.0); D(98.5, 105.0, 200.0); E(130.0, 423.5, 255.0); F(370.0, 423.5, 255.0); G(370.0, 273.5, 255.0); and H(130.0, 273.5, 255.0). The room was divided into 50 areas, each 50 × 50 cm^2, but only 29 areas, suitably distributed and covering the entire room, were considered in our analyses. The array positions and areas are depicted in Fig. 7.

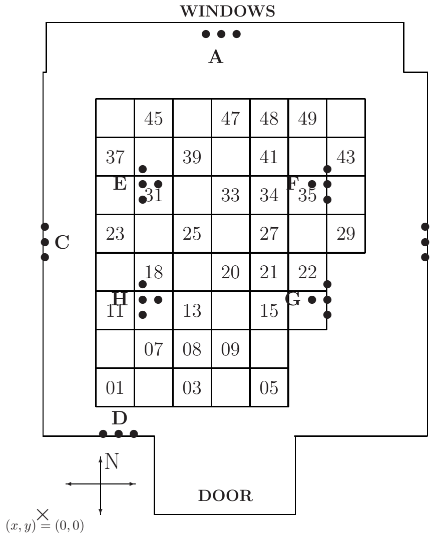

FIG. 7. The environmental setup used in the automatic estimation experiment. The areas considered in the study are numbered.

A loudspeaker was set-up over a stand fixed 140 cm above the floor to simulate an acoustic source. The stand was put in the center of each numbered area indicated in Fig. 7, and 300 Japanese words from two adult male speakers were played, with the average duration of an utterance being 0.6 s. In each studied area, the loudspeaker was turned to reflect 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) and 34 800 samples (29 × 4 × 300). Utterances were recorded at 48 kHz by a 32 channel acquisition system and downsampled to 12 kHz. The acquisition and the playback equipment are separated, each one running on different computers with Linux operating system. The acquisition board is manufactured by Tokyo Electron Device Limited, the loudspeaker is an Entry S type manufactured by ALR Jordan, and the microphones are ECM-C10 model produced by Sony Corporation. The small loudspeaker emits the sound energy by its front part and attenuates at its rear. Thus, in this work we assumed that the loudspeaker could roughly model the human head. The background noise was around 35 dBA measured by the Ono Sokki (model LA-4350) sound meter. The estimated SNR estimated by the recorded signals was 15 dB, sufficient to reduce the performance of the GCC-PHAT estimation method when the estimates were determined by frames, which was verified experimentally. In the GCC-PHAT analysis, a frame length of 256 samples, frame shift of 128

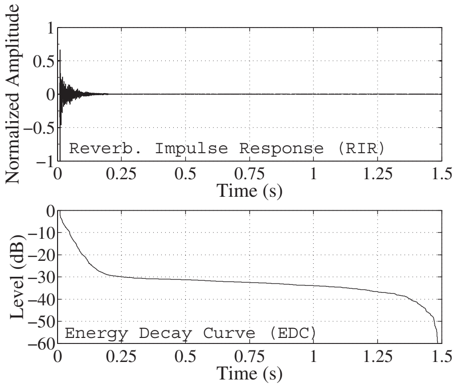

FIG. 8. RIR and EDC measured from the center of area 20 (see Fig. 7) to the center microphone of array A.

samples, and Hamming window were considered. In the experiments, one source position estimate was obtained per utterance. The reverberation impulse response (RIR) and the energy decay curve (EDC) are measured from the center of position 20 (see Fig. 7) to the center microphone in array A when a loudspeaker was turned to the north. The RIR and EDC are illustrated in Fig. 8. The reverberation time T_{60} corresponds to the time lag when the EDC decays 60 dB; here it is longer than 1 s.

B. Automatic estimation: Baselines for position estimation

Tables I and II present the results of the baseline position estimation methods in terms of the average position error at each array (arrays A, B, C, D, E, F, G, and H) in 2D/3D spaces. An analysis of Table I shows that the SRP-PHAT$_{\text{array}}$ is more robust than the TDOA-based method, and arrays fixed at the ceiling (arrays E, F, G, and H) yield better results than arrays fixed at the walls (arrays A, B, C, and D). In Table II, the *oracle* selection is performed for the best array selection for the TDOA-based and the SRP-PHAT$_{\text{array}}$ methods, i.e., the array with the best source position estimate is always selected. In the SRP-PHAT$_{\text{all}}$ method, in which one source position estimate is obtained by the entire array network, we had the best baseline results.

C. Automatic estimation: Two-stage ANNs for position and orientation estimation

ANNs were studied using the Stuttgart neural network simulator (SNNS).[27] For the ANN training/testing phase, the recorded data from each of the 29 areas were divided into two complementary sets with 80% of the data used in the training phase and 20% in the testing phase and with no overlap between both data sets. In the results, CLOSED and OPEN tests refer to the results obtained by the trained ANN evaluated using the training and testing data sets, respectively. Cross-validation was performed to generalize the results of the proposed position and orientation estimation method. This process was performed five times, but permuting the original data set. In the experiments, the average position error was calculated in 2D (x,y) and 3D (x,y,z) spaces. The following measures were used to express the results in the evaluation phase of the ANNs: the correct orientation ratio (%), which expresses the agreement between the estimated and the true source orientation, the average orientation error (°), which denotes the angle mismatch between the estimated and the true source orientation, the best array selection ratio (%), which expresses the agreement between the chosen array and the array nearest to the source position, and the average position error (centimeters), which denotes the mismatch between the estimated and the true source position. In the ANN training phase, a sufficient number of iterations assures convergence to a steady state or to avoid overtraining.

TABLE II. The average position error of the baselines: the TDOA-based, SRP-PHAT$_{\text{array}}$, and SRP-PHAT$_{\text{all}}$.

	Space	
Method	2D (cm)	3D (cm)
TDOA-based (best array—oracle)	21.2	34.1
SRP-PHAT$_{\text{array}}$ (best array—oracle)	18.6	31.7
SRP-PHAT$_{\text{all}}$	16.0	29.8

Figure 6 illustrates the gating two-stage ANN, and Table III presents the studied configurations. In the training phase, at the first stage, the correct orientation was used as the target value, while at the second stage either the best array or the actual source position was used. The ANN$_1$ inputs consist of the combination of the sets of energy features and distance estimates {P+C+D}, which estimates the orientation at the first stage (one out of four output values) and selects the best array at the second stage (one out of eight output values). The ANN$_2$ inputs consist of the combination of the sets of energy features and source position estimates $\{\text{P}+\text{C}+\{\hat{x},\hat{y}\}/\{\hat{x},\hat{y},\hat{z}\}\}$, which estimates the orientation in the first stage and the position in 2D/3D space in the second stage. Finally, the ANN$_3$ inputs consist of the combination of the sets of energy features, microphone positions, and TDEs $\{\text{P}+\text{C}+[x,y,z]+\{\hat{\boldsymbol{\tau}}\}\}$, which estimates the orientation in the first stage and the position in 3D space in the second stage.

TABLE I. The average position error by array considering the baseline TDOA-based and SRP-PHAT$_{\text{array}}$ estimation methods. The position of arrays {A, B, C, D, E, F, G, and H} is defined in Sec. IV A and illustrated in Fig. 7.

	Average position error by array—2D/3D spaces (cm)							
Method	A	B	C	D	E	F	G	H
TDOA-based	659.9/678.2	396.6/407.6	372.4/384.4	585.4/600.6	160.0/207.8	191.1/242.7	223.9/270.2	154.2/209.4
SRP-PHAT$_{\text{array}}$	105.0/114.9	112.2/123.5	95.4/107.5	117.6/125.8	63.4/79.0	76.5/92.2	70.0/86.3	79.3/97.9

TABLE III. Two-stage ANN configurations. The results were obtained considering the hidden unit numbers in bold. (1) and (2) in output denote first and second stages, respectively.

	No. of ANN units		
Topology	Input	Hidden	Output
ANN_1	24 (P+C+D)	**48**	4(1),8(2)
ANN_2	32 (P+C+$\{\hat{x},\hat{y}\}$)	**80**	4(1),2(2)
ANN_2	40 (P+C+$\{\hat{x},\hat{y},\hat{z}\}$)	**80**	4(1),3(2)
ANN_3	120 ([x,y,z]+$\{\hat{\boldsymbol{\tau}}\}$)	**240**	4(1),3(2)
ANN_3	136 (P+C+[x,y,z]+$\{\hat{\boldsymbol{\tau}}\}$)	**272**	4(1),3(2)

The input set {P+C+[x,y]+$\{\hat{\boldsymbol{\tau}}\}$} is not presented here because the results were closed to the values obtained using the 3D results disregarding the z-dimension.

D. Automatic estimation: Orientation estimation using only energy features

In Ref. 20, the orientation of a directional acoustic source was estimated using 448 microphones which were used to estimate the radiation pattern of the source. Here, an experiment was evaluated to verify if the ANN can roughly model the source radiation pattern like of Sachar *et al.* using a small number of microphones and thus predict the source orientation. The set of energy features P and C, as defined in Sec. III A, was used in the orientation estimation. A simple one-stage fully connected feedforward ANN, like the first stage of the proposed two-stage ANN, illustrated in Fig. 6, with 16 input units {P+C}, 48 hidden units, and four (E, N, S, W) output units was studied. The number of hidden units was determined experimentally. The results of this experiment are discussed in Sec. IV E 3.

E. Automatic estimation: Results

1. Best array selection by individual criteria

In this section, an experiment was used to determine whether signal power, signal correlation, and distance estimate could be used separately to choose the best array. Three individual criteria were defined: the best array has maximum power estimate ($\mathcal{P}$), it has maximum correlation estimate ($\mathcal{C}$), and it has minimum distance estimate ($\mathcal{D}$) among all arrays. Each criterion must match the array which yields the minimum position error in a total of 34 800 (300 utterances × 4 orientations × 29 areas) input patterns. The results are in Table IV, where the correct selection ratio (L_1) is the relation between the number of matches and the total input patterns, and the average position error (L_2) is calculated by taking the selected arrays. Using SRP-PHAT$_{\mathrm{array}}$ position estimates we obtained better results in terms of L_2 than using TDOA-based position estimates, but these results are worse compared to baselines in Table II. Finally, it is clear by analyzing L_1 results that the selection of the best array by each individual criterion is not a suitable approach. This leads us to investigate a method, in our case exploring neural networks, to obtain more confident results.

TABLE IV. The correct selection ratio in percent (L_1) and the average position error in centimeters (L_2) when the minimum position error calculated in 2D/3D spaces matches the criteria defined in Sec. IV E 1. Source position estimates by the TDOA-based and SRP-PHAT$_{\mathrm{array}}$ methods were used to calculate position errors.

	TDOA-based		SRP-PHAT$_{\mathrm{array}}$	
Criteria	L_1	L_2	L_1	L_2
$\mathcal{P}$ and min. error in 2D	36.4	161.1	32.4	51.9
$\mathcal{C}$ and min. error in 2D	37.1	249.9	33.6	49.8
$\mathcal{D}$ and min. error in 2D	21.7	122.5	31.9	53.2
$\mathcal{P}$ and min. error in 3D	35.5	223.5	31.6	67.6
$\mathcal{C}$ and min. error in 3D	35.7	331.0	31.6	66.3
$\mathcal{D}$ and min. error in 3D	5.4	175.0	16.5	85.9

2. Proposed position and orientation estimation method

Table V presents the results of the proposed position and orientation estimation method, with configuration ANN_1 defined in Table III, using the sets of energy features P and C, and distance estimates D derived by TDOA-based and SRP-PHAT$_{\mathrm{array}}$ source position estimates, respectively. The results show that the SRP-PHAT$_{\mathrm{array}}$ has slightly better performance than the TDOA-based method in terms of the used measures. Table VI presents the results of the configuration ANN_2 defined in Table III using the set of energy features P and C and the set of source position estimates $\{\hat{x},\hat{y}\}$ or $\{\hat{x},\hat{y},\hat{z}\}$ obtained by the TDOA-based and SRP-PHAT$_{\mathrm{array}}$ methods, respectively. The results show that the SRP-PHAT$_{\mathrm{array}}$ is better than the TDOA-based method in terms of correct orientation ratio and average position error in this approach. Table VII presents the results of the configuration ANN_3 defined in Table III using the set of energy

TABLE V. Results obtained by the neural network topology ANN_1. The distance estimate set D was calculated by the TDOA-based (TDOA) and SRP-PHAT$_{\mathrm{array}}$ (SRP) source position estimates. The average position error value was calculated using the position estimates from the selected arrays for the given dimensional space.

			Corr. orient. ratio (%)		Avg. orient. error (deg)		Corr. select. ratio (%)		Avg. pos. error (cm)	
Inputs	Space	Test	TDOA	SRP	TDOA	SRP	TDOA	SRP	TDOA	SRP
24(P+C+D)	2D	CLOSED	80.7	81.7	23.8	22.5	60.3	59.2	30.9	27.3
24(P+C+D)	2D	OPEN	77.8	79.3	27.1	25.5	56.6	55.6	32.6	28.5
24(P+C+D)	3D	CLOSED	78.9	79.4	25.8	25.4	63.1	62.3	43.9	41.8
24(P+C+D)	3D	OPEN	76.0	76.8	29.3	28.6	59.2	58.7	45.8	43.4

TABLE VI. Results obtained by the neural network topology ANN_2. Source position estimates $\{\hat{x},\hat{y}\}/\{\hat{x},\hat{y},\hat{z}\}$ in 2D/3D were estimated by the TDOA-based (TDOA) and $\text{SRP-PHAT}_{\text{array}}$ (SRP) methods. The average position error value was calculated using the position estimates for the given dimensional space, and the value (*) was the average position error in 2D directly calculated from the 3D estimates disregarding the z-dimension.

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

features P and C, the set of TDEs $\{\hat{\boldsymbol{\tau}}\}$, and the set of microphone positions [x, y, z]. The columns marked by "△" assume the training and testing conditions in Sec. IV C. However, in the columns marked by "□," the training and testing conditions were modified. The modification consisted of separating training and testing data sets by areas. Six areas were randomly chosen inside the room, avoiding border areas. (Areas 1, 3, 5, 11, 22, 23, 29, 37, 43, 45, 47, and 49 are border areas in Fig. 7. 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. CLOSED and OPEN tests were performed using the training and testing data sets, and the data from all areas were common to both data sets in △ characterizing *CLOSED POSITONS TEST*, while in □ the data from areas in the training data set were not in the testing data set characterizing *OPEN POSITONS TEST*. Comparing results of Table VII to Tables V and VI, the best results in terms of correct orientation ratio and average position error were obtained, where *CLOSED POSITONS TEST* yielded better results than *OPEN POSITONS TEST*.

3. Orientation estimation by only energy features

In the CLOSED/OPEN test results of the one-stage ANN using only the set of energy features P and C, we had correct orientation ratios of 56.2%/57.4%, respectively, and even increasing the number of iteration in the ANN training phase, there was not a significant improvement in the ratio. This behavior seemed to be strange since it was expected that the ANN could model the source radiation pattern in order to predict the source orientation. A possible cause of this divergence could be due to small amount of energy feature samples, which were restricted to 16 samples {P+C}. The orientation estimation task by using only energy features showed to be difficult even using the features of all array of the array network.

V. DISCUSSION

A. Two-stage ANNs for position and orientation estimation

As shown in Table V, the set of distance estimates D derived from the TDOA-based source position estimates yielded better results in terms of the best array selection than using the $\text{SRP-PHAT}_{\text{array}}$ source position estimates. However, for the correct orientation ratio and average position error, the opposite was observed, for instance, in the OPEN results in terms of the correct orientation ratio, the best array selection, and the average position error were 77.8% (76.0%), 56.6% (59.2%), and 32.6 cm (45.8 cm) in 2D (3D) space for the TDOA-based method and were 79.3% (76.8%), 55.6% (58.7%), and 28.5 cm (43.4 cm) in 2D (3D) space for the $\text{SRP-PHAT}_{\text{array}}$ method. The results in terms of the average position error were worse than the baseline presented in Table II, but encourage us to continue our investigation seeking for better results.

From Table VI, it is evident that $\text{SRP-PHAT}_{\text{array}}$ source position estimates yielded better results than the TDOA-

TABLE VII. Results obtained by the neural network topology ANN_3. TDEs and microphone positions of every array were employed. "△" denotes *CLOSE POSITONS TEST* and "□" denotes *OPEN POSITONS TEST*. The average position error value was calculated using the position estimates for the given dimensional space, and the value (*) was the average position error in 2D directly calculated from the 3D estimates disregarding the z-dimension.

		Corr. orient. ratio (%)		Avg. orient. error (deg)		Avg. pos. error (cm)	
Inputs	Test	△	□	△	□	△	□
$120([x,y,z]+\{\hat{\boldsymbol{\tau}}\})$	CLOSED	99.9	96.8	0.1	4.2	24.4 (20.9*)	24.9 (21.6*)
$120([x,y,z]+\{\hat{\boldsymbol{\tau}}\})$	OPEN	99.4	87.0	0.7	17.0	24.6 (21.0*)	25.3 (22.5*)
$136(\text{P}+\text{C}+[x,y,z]+\{\hat{\boldsymbol{\tau}}\})$	CLOSED	99.9	96.3	0.1	4.5	23.2 (20.3*)	25.7 (22.1*)
$136(\text{P}+\text{C}+[x,y,z]+\{\hat{\boldsymbol{\tau}}\})$	OPEN	99.5	84.2	0.6	18.5	23.3 (20.5*)	28.2 (24.8*)

based, for instance, in OPEN results, the correct orientation ratio and the average position error were 83.9% (90.3%) and 31.1 cm (33.9 cm) in 2D (3D) space for TDOA-based method and were 94.7% (98.3%) and 23.5 cm (27.9 cm) in 2D (3D) spaces for SRP-PHAT$_{\text{array}}$ method. Comparing Tables V and VI, the improvement in the orientation estimation changing the set of distance estimates D by the set of source position estimates $\{\hat{x},\hat{y}\}/\{\hat{x},\hat{y},\hat{z}\}$ can be explained by the fact that a 2D/3D coordinate has more spatial information than a one-dimensional value defined by D. The ANN$_2$ configuration yielded a correct orientation ratio of 90.3% using source position estimates by TDOA-based and 98.3% using source position estimates by SRP-PHAT$_{\text{array}}$, while the ANN$_1$ configuration yielded a correct orientation ratio of 76.0% using distance estimates derived from TDOA-based and 76.8% using distance estimates derived from SRP-PHAT$_{\text{array}}$ for OPEN results in 3D space. In Table VI, the source position estimates by SRP-PHAT$_{\text{array}}$ yielded better results (27.9 cm) than the baseline SRP-PHAT$_{\text{all}}$ (29.8 cm) for the 3D case in terms of the average position error, but for the 2D case, the SRP-PHAT$_{\text{all}}$ is still the best. Values (*), in Table VI, are the 2D estimates obtained directly by the 3D estimates disregarding the *z* dimension, which showed to be better estimates than those obtained by training the ANNs using 2D source position estimates.

Comparing the results in Table VII with those in Tables V and VI, we noted that the set of TDEs and the set of microphone positions (low level parameters) were more suitable for the orientation and position estimation than the set of source position estimates (intermediate level parameters) or the set of distance estimates (high level parameters). In *CLOSED POSITONS TEST* in Table VII, a correct orientation ratio of 99.5% and an average position error of 23.3 cm in OPEN results in 3D case were obtained, a reduction of 21.8% compared to the baseline SRP-PHAT$_{\text{all}}$. However, SRP-PHAT$_{\text{all}}$ (16.0 cm) is still better in 2D case compared to the estimation method (20.5 cm). In the *OPEN POSITONS TEST*, the objective was to jointly estimate the position and orientation in the untrained areas. The results show that the correct orientation ratio ranges from 84% to 87% and the average position error is 25.3 cm, which is still better than the baseline SRP-PHAT$_{\text{all}}$ in 3D space.

In *OPEN POSITONS TEST*, adding the set of energy features P and C in the input data set, we had an increase in the reliability of the positions used in the training phase. Thus, the ANNs were more dependent on these positions. In the testing phase, the input pattern tends to get closer to the positions used in the training phase, and the orientation is affected by this estimation because of overtraining. Therefore, there is an increase in the position error and decrease in the orientation ratio due to the confidence given by the set of energy features in the trained areas, which reduce the performance in the untrained areas in *OPEN POSITONS TEST*, as observed in Table VII.

B. Additional comments: Automatic estimation

For different representation levels of the estimated parameters, an improvement in the results and in the number of input units in the ANN was observed by moving from the high to low level of representation, which can be observed by comparing the results of Tables VII (using high representation), V (using intermediate representation), and VI (using low representation). Although the analyses presented in this paper concentrate on using all eight arrays, other analyses could be performed by individual array or combining different arrays, such as using only walls or ceiling arrays. Ceiling arrays are especially important because they can perceive signals directly and it is more difficult physically to block them than wall arrays. For instance, considering energy features P and C, TDEs $\{\hat{\boldsymbol{\tau}}\}$, and microphone positions $[x,y,z]$ sets in the experiments, the correct orientation ratio (%) and average position error (2D/3D) by using only array **A** (wall array) were 67.9% and 48.7/53.0 cm, using only array **E** (the ceiling array yielded the best average position error in Table I) were 76.0% and 38.6/45.4 cm, and the combination of arrays **E** and **G** (ceiling arrays) was 88.5% and 26.4/30.7 cm in training and testing conditions of Sec. IV C.

The estimated position and orientation by the proposed method could be used in some practical application, for example, in automatic speech or speaker recognition. In Ref. 6, the source position estimate was used to compensate the microphone signals due to the channel effect. It was shown that in a squared region of $60\times 60\ \text{cm}^2$ around the sound source there was no much effect in recognition results; thus the estimated position by our proposed method could be used in recognition tasks. The estimated orientation could be used in a compensation method too. As a directional sound source does not radiate equally in all directions, suitable weights can be applied to microphone signals that are not in front of the speaker, compensating attenuation and distortion due to the relative orientation prior to a signal processing step.

C. Additional comments: Could the proposed automatic estimation method model the human auditory perception system?

Here, a short discussion about the proposed method and the human auditory perception system is presented. A subjective experiment with blindfolded listeners was performed in the same environment. After hearing a spoken phrase by a human speaker, the listener was allowed to remove the blindfold and indicate the speaker facing angle (E, N, S, or W) and the speaker position inside the room. Six listeners participated in the experiment. The results showed correct orientation ratios of 76.7% and 79.9% with average position errors of 65.4 and 60.5 cm in 2D (Ref. 28) before and after a training phase, respectively. In the training phase, the listener was allowed to remove the blindfold and verify the speaker position and orientation. Comparing these results with the best results of the proposed automatic method, we noted that the proposed method using the entire microphone array network outperforms the human auditory perception ability to discriminate the position and orientation of an acoustic source. However, it is unfair to compare the array network, which is distributed in the space, with the auditory system, which is concentrated in the head. A more fair comparison would be with an individual array. It is clear that arrays **A** (wall array near the window) and **E** (ceiling array), individually, do not have the same performance as the array network, as noted by the results in Sec. V B. The average position errors obtained by individual arrays **A** and **E** were better than those obtained in the subjective experiment, but in terms of the correct orientation ratio the auditory system yielded better estimates. Thus, one array cannot model the auditory perception system accurately in terms of orientation estimation. Additionally, exploring the spatial diversity by combining two ceiling arrays **E** and **G** showed to be sufficient to surpass the auditory perception system in terms of the considered measures.

VI. CONCLUSIONS

In this paper, a method which automatically estimates the position and orientation of a directional acoustic source by neural networks was proposed. Different strategies for the automatic estimation changing the type of the input parameters were compared. The research was restricted to fully connected feedforward neural networks, but other ANN types and topologies could also be tested. The proposed method using low level parameter (TDEs and microphone positions) input set yielded better results than intermediate (source position estimates) and high parameter (distance values) input sets.

In terms of position estimation, the proposed method was better than the conventional position estimation methods (TDOA-based and SRP-PHAT) presented in this paper at least in 3D space. In terms of correct orientation ratio, accurate estimation was obtained using low or intermediate level parameters. As the position information is important in speech recognition, it is expected that employing the orientation information will be useful in this task, for instance, either selecting the nearest array facing the speaker or restricting the vocabulary that must be recognized in a given direction in a voiced command application, supposing that the speaker faces the device before uttering the command. Finally, as presented in last examples, speech recognition application will be in our next research step.

ACKNOWLEDGMENTS

We would like to thank the Global COE program "Frontiers of Intelligent Sensing" and MEXT for supporting our research.

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

[2] E. Hansler and G. Schmidt, *Speech and Audio Processing in Adverse Environments* (Springer, New York, 2008).

[3] J. G. Ryan and R. A. Goubran, "Optimum near-field performance of microphone arrays subject to a far-field beampattern constraint," J. Acoust. Soc. Am. **108**, 2248–2255 (2000).

[4] J. C. Chen, R. E. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field," IEEE Trans. Signal Process. **50**, 1843–1854 (2002).

[5] X. Chen, Y. Shi, and W. Jiang, "Speaker tracking and identifying based on indoor localization system and microphone array," in 21st International Conference on Advanced Information Networking and Applications Workshops (2007), Vol. **2**, pp. 347–352.

[6] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distance speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM," Speech Commun. **49**, 501–513 (2007).

[7] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in video conferencing," in Proceedings of the ICASSP (1997), Vol. **I**, pp. 187–190.

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

[9] M. R. Bai and C. Lin, "Microphone array signal processing with application in three-dimensional spatial hearing," J. Acoust. Soc. Am. **117**, 2112–2121 (2005).

[10] 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 the ICASSSP (2006), Vol. **IV**, pp. 929–932.

[11] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," J. Acoust. Soc. Am. **107**, 384–391 (2000).

[12] R. Bucher and D. Misra, "A synthesizable vhdl model of the exact solution for three-dimensional hyperbolic positioning system," VLSI Des. **15**, 507–520 (2002).

[13] A. Brutti, M. Omologo, and P. Svaizer, "Speaker localization based on oriented global coherence field," Proceedings of the Interspeech (2006), pp. 2606–2609.

[14] J. Chen, J. Benesty, and Y. Huang, "Performance of gcc-and amdf-based time-delay estimation in practical reverberant environment," EURASIP J. Appl. Signal Process. **1**, 25–36 (2005).

[15] F. Talantzis, A. G. Constantinides, and L. C. Polymenakos, "Estimation of direction of arrival using information theory," IEEE Signal Process. Lett. **12**, 561–564 (2005).

[16] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-24**, 320–327 (1976).

[17] V. C. Raykar, "Automatic position calibration of multiple microphones," in Proceedings of the ICASSSP (2004), Vol. **IV**, pp. 69–72.

[18] I. McCowan, M. Lincoln, and I. Himavan, "Microphone array shape calibration in diffuse noise fields," IEEE Trans. Audio, Speech, Lang. Process. **16**, 666–670 (2008).

[19] A. Brutti, M. Omologo, and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart room equipped with distributed microphone arrays," in Proceedings of the Interspeech (2005), pp. 2337–2340.

[20] J. M. Sachar and H. F. Siverman, "A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array," in Proceedings of the ICASSSP (2004), Vol. **IV**, pp. 65–68.

[21] A. Abad, C. Segura, C. Nadeu, and J. Hernando, "Audio-based approaches to head orientation estimation in a smart-room," in Proceedings of the Interspeech (2007), pp. 590–593.

[22] C. Segura, C. Canton-Ferrer, A. Abad, J. R. Casas, and J. Hernando, "Multimodal head orientation towards attention tracking in smartrooms," Proceedings of the ICASSP (2007), Vol. **II**, pp. 681–684.

[23] A. Brutti, M. Omologo, P. Svaizer, and C. Zieger, "Classification of acoustic maps to speaker position and orientation from a distributed microphone network," in Proceedings of the ICASSSP (2007), Vol. **IV**, pp. 493–496.

[24] S. Haykin, *Neural Networks: A Comprehensive Foundation* (Macmillan, New York, 1994).

[25] K. Varma, T. Ikuma, and A. A. Beex, "Robust tde-based doa estimation for compact audio arrays," in 2nd IEEE Sensor Array and Multichannel Signal Processing Workshop (2002), pp. 214–218.

[26] R. Parisi, A. Cirrillo, M. Panella, and A. Uncini, "Source localization in reverberant environments by consistent peak selection," in Proceedings of the ICASSSP (2007), Vol. **I**, pp. 37–40.

[27] http://www.ra.cs.uni-tuebingen.de/SNNS/ (Last viewed 3/24/2009).

[28] A. Y. Nakano, K. Yamamoto, and S. Nakagawa, "Auditory perception of speaker's position, distance and facing angle in a real enclosed environment," in Acoustical Society of Japan—Autumn Meeting (2008), pp. 525–526.