Estimation of user orientation using GMMs for multiple voice-commanded devices environments

Alberto Yoshihiro Nakano
UTFPR - Campus Toledo
Rua Cristo Rei, 19 - 85902-490
Toledo - PR - Brazil
nakano@utfpr.edu.br

Phillip Mark Seymour Burt
Dept. Telecommunications and Control - Escola Politécnica -USP
Av. Prof. Luciano Gualberto, tr. 3, 158 - 05508-900
Cidade Universitária - São Paulo - SP - Brazil
phillip@lcs.poli.us.br

***Abstract*— A friendly interface for electronic devices has become more important in our lives, mainly, for elderly and disabled people. In this sense, devices equipped with an automatic speech recognition (ASR) system, that is, voice-commanded devices, can properly address this issue, however, we still would need to manually input information when noisy and reverberant conditions were considered. When dealing with multiple voice-commanded devices in the same environment, additional information about the speaker, such as his position, would be beneficial to improve the ASR performance by using, for instance, beamforming techniques to enhance the signal of interest. In this work, we study the orientation of a loudspeaker, modeling a human speaker, detected by a pair of two microphones, as additional information which would complement the position information in the previous situation. We show that binaural cues used as inputs for Gaussian mixture models (GMMs) can be used to discriminate an orientation among a defined set of orientations.**

***Index Terms*— ASR, binaural cues, microphone array, source orientation estimation.**

I. Introduction

In recent decades, advances in technology have gradually changed the way we live, improving our lifestyle but making us more dependent on it. The evolution of technology has increasingly required technical knowledge from the user, who often is not prepared to deal with it. For instance, the increased number of functions and buttons on remote controls may confuse, rather than help the user. Thus, an easy way to access a system is of fundamental importance. Access should be simple, easy, and intuitive, such as a plug-and-play system, avoiding complex procedures.

ASR system is an option to be considered for interfaces, because speech is the most common form of interaction between humans. This mechanism of interaction would make access to the system simpler, dynamic, and efficient. However, when either environmental conditions are unfavorable or training conditions are different from the test conditions in the ASR system, the system tends to fail. In this case, we still would need to manually input information. To cope with environmental conditions, beamforming [1], [2], noise reduction [3], [4], speech enhancement [5], [6], [7], and speech dereverberation [8], [9] can be used in order to improve the quality of the signals used in the recognizer. The mismatch between training and test conditions in an ASR system can be handled through feature-based compensation or model-based adaptation techniques.

Currently, information about the user, such as its estimated location using microphone arrays has been treated as additional information for ASR systems, however little attention has been given to the orientation of the user. In this work we show that the user's orientation can complement the user's position when dealing with multiple voice-commanded devices.

The estimation of the head orientation (source orientation) [10] of a user has gained interest with applications such as intelligent robots [11], voice-commanded devices in a smart room [12], [13], and control of powered wheelchairs for disabled people [14], among others. In most cases, estimation of the source orientation is based on a model of the directivity of the source and on the intensity of the signals measured at each microphone, although in [15] we verified that taking also time delays of microphone pairs into account led to a significant increase in the precision of the estimate.

Specifically in smart environments, how would the user deal with multiple voice-commanded devices without making access mistakes? In [13], artificial neural networks (ANNs) were used to estimate the orientation of the user, assuming that the user turns to the device he wants to control, and then applied to discriminate different controlled devices. In that approach a distributed microphone array was used to deal with the spatial diversity.

The use of a large array [13], [16] may provide robustness against unaccounted contributions to signal intensity, such as microphone directivity and reverberation. However, inspired by the human auditory system, that can roughly estimate the orientation of an out-of-sight acoustic source, we consider an array of only two microphones and investigate the use of the binaural cues interaural time difference (ITD) and interaural level difference (ILD), for orientation estimation. This is done by maximizing the likelihood of a Gaussian mixture model (GMM) created for each discrete orientation. GMM was chosen because it is a well-known method used in speaker and speech recognition tasks, whose framework can be easily modified for the orientation estimation task.

The paper is organized as follows. Section II discusses the problem encountered in an environment with multiple voice-

commanded devices. Section III presents the modeling of the user and the device interface by a loudspeaker and an array of two microphones. In Section IV the proposed GMM-based orientation estimation method is presented. Section V presents the experimental setup. Finally, Sections VI and VII present the results and the final conclusions, respectively.

II. Problem description

A friendly interface to access electronic devices must be reconfigurable, reliable, appealing and simple to use. It should also be intuitive to use mainly for elderly, handicapped, and disabled people. In this sense, ASR system embedded devices can properly address this problem by allowing the execution of predefined tasks through spoken commands. This "tell-execute" interaction describes an intuitive way to operate a device. Classical interfaces such as remote controls will still be available to manually input information when either noisy and reverberant conditions are found or a system configuration is required, providing different interaction options to the user. However, even when environmental conditions are favorable, in an environment equipped with voice-commanded devices, as shown in Figure 1 (top), the simple act of controlling a particular device such as a TV can become a difficult task. When the user commands one device to turn on, other devices may turn on too. This situation is showed in Figure 1 (middle). To avoid such a problem, a mechanism to correctly select the desired device is needed. Imagine a new situation, illustrated in Figure 1(bottom), in which the user looks at the device he wants to control, and in this case, after a spoken command, only one device is activated. Thus, exploring the user orientation information, we would restrict user access to just one device at a time.

III. User/speaker and device modeling

In order to employ the user's orientation as an additional information to select one device, a loudspeaker placed on a rotating table and an array of microphones were used to model the user and the device interface, respectively, as illustrated in Figure 2.

A Roland DS-7 loudspeaker was taken as the directional acoustic source modeling the user's head. For the sake of simplicity, we assume that the normalized directivity of the source is given by

$$D(\phi, k, J) = \frac{k + \cos^{2J}(\phi/2)}{k+1}, \qquad (1)$$

where ϕ is the azimuth with respect to the direction the source is aiming, k and J control the directivity ($J = 0$ leads to an omnidirectional pattern whereas $J > 0$ leads to a directional pattern; $k > 0$ guarantees $D(\phi = 180°, k, J) > 0$). This model improves the basic cardioid-like emission pattern, like the one presented in [17], by including the energy irradiated backward, controlled by k. In Figure 3, the measured directivity of the loudspeaker used in the experiments is compared to the model with $k = 0.05$ and $J = 3$.

Regarding the device we aim to control, it was assumed that its interface has microphones for signal acquisition. Aiming at greater practical interest and inspired by the human auditory system, an array of only two omnidirectional microphones was considered in this study. The playback and recording procedures for data acquisition are presented in detail in Section V.

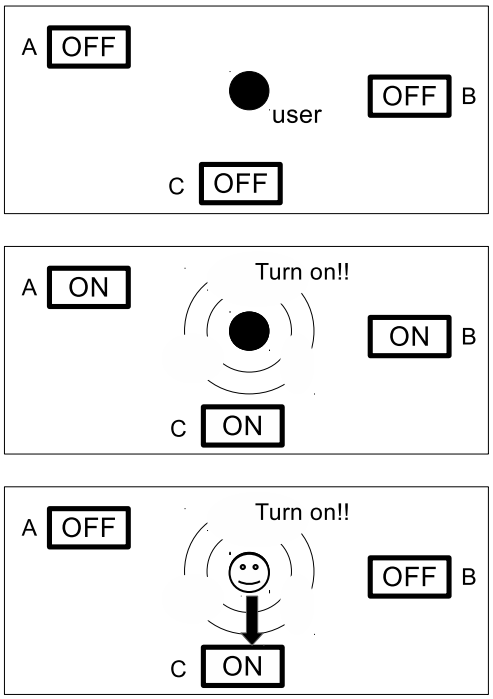

Fig. 1. (Top): smart environment with three ASR system embedded devices A, B, and C; (Middle): failed attempt to turn on only the device C; (Bottom): success to turn on only the device C.

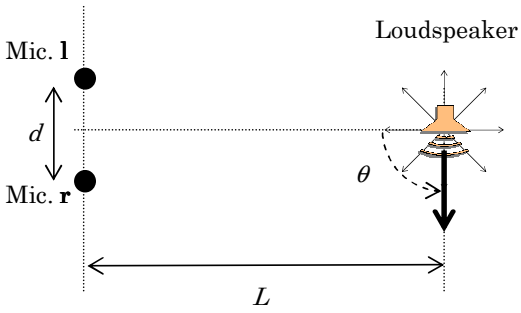

Fig. 2. Loudspeaker and array (left and right microphones) used to model the user and the device interface. d is the distance between microphones, L is the distance between the loudspeaker and the array, and θ is the loudspeaker orientation angle.

A. Binaural cues: ITD and ILD

The difference in arrival time and the difference in intensity of a sound at two ears are defined as ITD and ILD, respectively. These two binaural cues are important for human perception of location and motion of acoustic sources. Given the microphone signals and assuming that the array

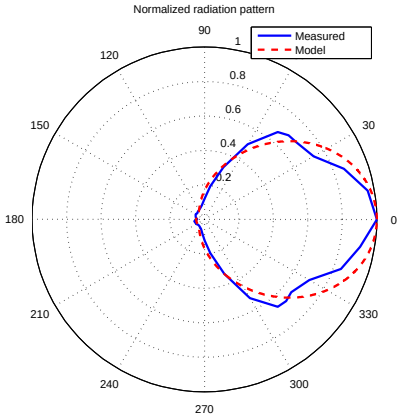

Fig. 3. Model with $k = 0.05$ and $J = 3$ (dashed line) and measured (continuous line) values of loudspeaker directivity.

can roughly model the human auditory system, the ILD is estimated by the ratio of the signal energies and the ITD is estimated using the generalized cross-correlation with phase transform (GCC-PHAT) function [18], [19],

$$R(\tau_{lr}) = \int_{-\infty}^{+\infty} \frac{X_l(f)X_r^*(f)}{|X_l(f)X_r^*(f)|} e^{-j2\pi f \tau_{lr}} df, \tag{2}$$

where $*$ represents conjugation, and $X_l(f)$ and $X_r(f)$ are the spectral representations of the binaural signals $x_l(t)$ and $x_r(t)$, respectively. The estimate of ITD, $\hat{\tau}_{lr}$, corresponds to the time difference that maximizes $R(\tau_{lr})$, as

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

Figure 4 illustrates the framework for binaural cues estimation, which could be easily integrated in an ASR system framework. $s(t)$ denotes the clean speech sample, $h_l(t)$ and $h_r(t)$ denote the measured binaural impulse response, subscript l and r denote the left and right microphones in the array, and $\hat{x}_l(t)$ and $\hat{x}_r(t)$ denote the signals measured at each microphone.

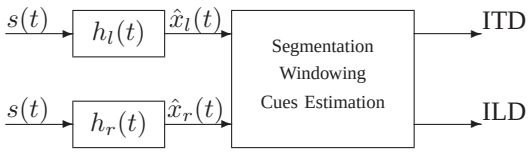

Fig. 4. Estimation of the binaural cues ITD and ILD.

B. Histograms

Figure 5 presents the histograms of the ITD and ILD for $\theta = \{0°, 90°, 270°\}$. The histograms reflect the variation of the statistical distribution of the binaural cues with orientation θ, defined in Figure 2. The variation of the mean value of the ILD with orientation can be explained by the directivity of the source, while the variation of the mean value of the ITD can be explained by the point sound source (the user's mouth or the frontend of the loudspeaker) and its center of rotation (the user's neck or the center of the rotating table) not being at the same point in space. These variations are explored in the GMM-based orientation estimation method.

IV. GMM-BASED ORIENTATION ESTIMATION

A. Gaussian Mixture Model

A GMM is the weighted sum of M component Gaussian densities expressed by [20]

$$P(\mathbf{x}|\lambda) = \sum_{i=1}^{M} c_i b_i(\mathbf{x}), \tag{4}$$

where $\mathbf{x}$ is a $\mathcal{D}$-dimensional data vector; $\lambda = \{c_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ denotes the GMM parametric model with mixture weights c_i, mean vector $\boldsymbol{\mu}_i$, and covariance matrix $\boldsymbol{\Sigma}_i$; and

$$b_i(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{5}$$

are the component densities, for $i = 1, \ldots, M$. Each component density is a $\mathcal{D}$-variate Gaussian function. The mixture weight is constrained by

$$\sum_{i=1}^{M} c_i = 1. \tag{6}$$

The dimensionality $\mathcal{D}$ of vector $\mathbf{x}$ vary depends on the number of features, for example, when modeling only ITD it has dimensionality one, when modeling ITD and ITD together it has dimensionality two, and so on.

B. Estimation Method

In Section III-B, histograms show that the statistical distributions of ITD and ILD vary with the user's orientation. Thus, creating specific statistical models for a discrete set of orientations could be used to obtain an estimate of orientation from measured values of ITD and ILD. The procedure is similar to that used for speaker recognition, where each speaker has a statistical model trained from samples of signals generated by the speaker himself. For a test signal, the speaker with the model that yields the highest likelihood is accepted as the speaker who generated the test signal.

Let $P(\mathbf{x}|\lambda(\theta))$ be the GMM created for orientation θ. Given the data set $\mathbf{x}$, the orientation that generated it is assumed as the one which maximizes the likelihood

$$\hat{\theta} = \arg\max_{\theta} \{P(\mathbf{x}|\lambda(\theta))\}. \tag{7}$$

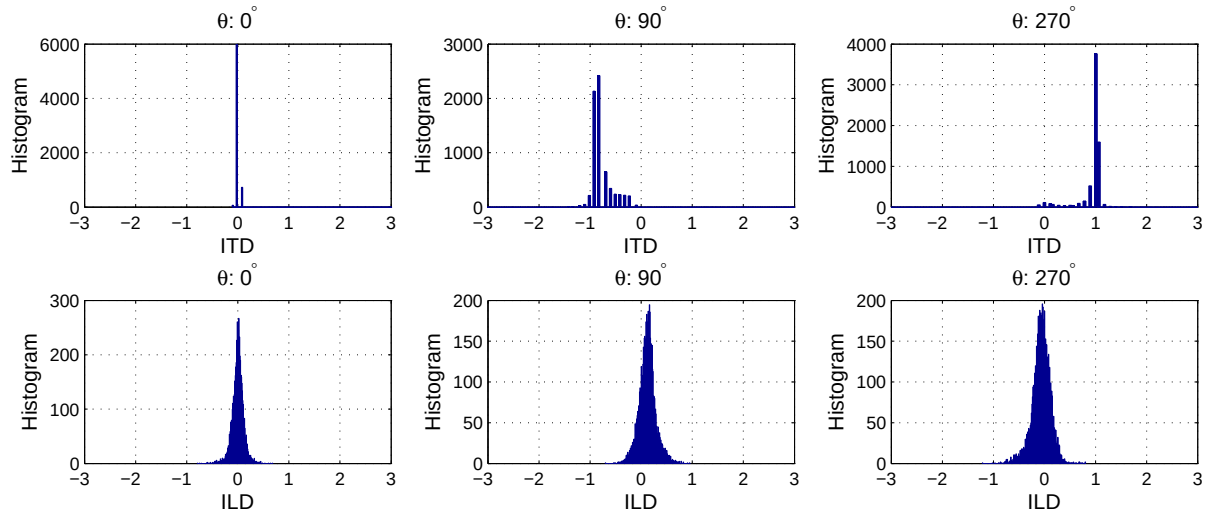

Fig. 5. Histograms of ITD and ILD for $\theta = 0°$ (the loudspeaker faces the array), $90°$, and $270°$.

V. Experimental Setup

The experiments were conducted in a soundproof room 3.0 m long, 2.7 m wide, and 3.0 m high. Its reverberation time T_{60} was around 130 ms. The measured background noise was lower than 30 dB (A-weighted). A Roland DS-7 loudspeaker was placed on a turntable with 360° of freedom and 1 m above the floor. The distance between the rotation axis of the table and the front of the loudspeaker was 13 cm. The array consisted of two omnidirectional Le Son microphones separated by $d = 13.6$ cm, 1 m above the floor, and $L = 1$ m away from the loudspeaker. Figure 6 illustrates the positions in the experimental setup. The A/D and D/A converters (Edirol FA-101) operating at 48 kHz sampling frequency was used for playback and recording.

For each of 8 loudspeaker orientations, a binaural impulse response was measured, downsampled to 16 kHz, and then convolved with a set of samples from the TIMIT database (two samples for each of 5 male and 5 female speakers). Using a frame length of 256 samples, a frame shift of 128 samples, and Hamming windowing, frame ITDs were obtained from the generalized cross-correlation with phase transform (GCC-PHAT) function and an interpolation method, for greater precision. Frame ILDs were taken as the ratio of the frame energies. All processing was executed using Matlab. GMMs were estimated using the Expectation-Maximization (EM) algorithm applied to the features extracted from the TIMIT data samples with full covariance matrices. Another data set was created using different samples from the same speakers from TIMIT and used as test set.

Two measures were used to evaluate the performance of the GMM-based system, the correct orientation ratio (COR), which expresses the agreement between the estimated and the true orientations, and the average orientation error (AOE), which denotes the angle mismatch between the estimated and the true orientations.

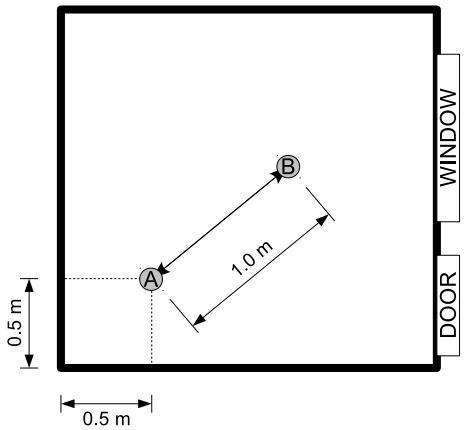

Fig. 6. Positions of the loudspeaker (A) and array (B) in the soundproof room.

VI. Results

A. Single frame estimation

Figure 7 shows the variation in the COR by increasing the number of Gaussian densities in the models. The behavior of the COR for ILD and ITD models was expected from the histograms presented in Figure 5. On the one hand, the similarities between the distributions of ILD for different orientations create a confusion in the adopted maximum likelihood strategy, resulting in a poor performance of the system. On the other hand, the dissimilar distributions of ITD for each orientation positively affect the decision making.

Combining the measured features allows us to improve the system performance as can be seen in Figure 7 in the ITD+ILD, E+ITD, and E+ILD+ITD curves, where E represents the pairwise frame energies that are readily available in the system framework. It is expected that E be greater when the loudspeaker is facing the array than when it has its back turned to the array. Thus, using E as a feature

could allow us to deal with the ambiguity that results from the use of only two microphones (for instance, 0° and 180° have the same ITD and ILD). It should be noted, however, that in the previous example they have the same theoretical value but different distributions, which, when combined, could improve the system performance. This fact can be observed in ITD+ILD compared to ILD or ITD in Figure 7.

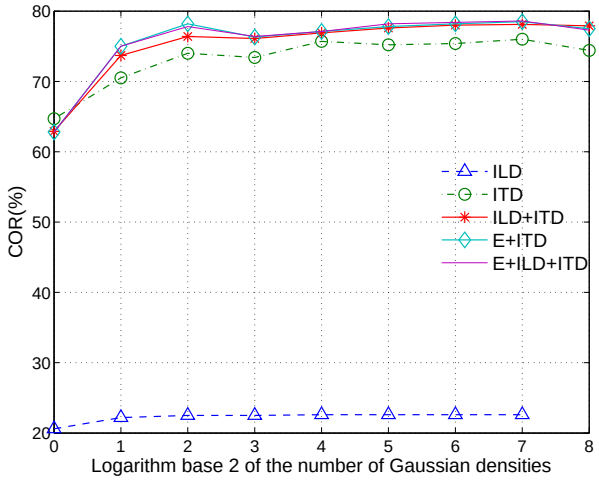

Fig. 7. Evaluation of the system in terms of COR (%) and the logarithm base 2 of the number of Gaussian densities $\{1, 2, 4, 8, 16, 32, 64, 128, 256\}$.

Considering GMMs with 32 Gaussian densities in Table I, the COR and AOE are presented according to the defined set of orientations. Note that we excluded the ILD results in the table due to its poor performance. As can be seen in COR columns in Table I, due to the angle ambiguity there is a lower performance of the system in the interval $135° \leq \theta \leq 225°$ in all cases.

B. Multiframe estimation

In a voice command system, the entire command utterance could be used in a multiframe estimation procedure, providing a tradeoff between COR and estimation delay. A simple multiframe estimation procedure is a majority voter, which, assuming independent frame estimations would have a multiframe COR_N given by

$$\text{COR}_N = 100 \times \sum_{k=0}^{N} P_k R_k, \tag{8}$$

where P_k is the probability that the correct orientation is chosen in k frames out of a total of N frames and R_k is the probability that in the remaining $N-k$ frames no other (incorrect) orientation is chosen in $k' \geq k$ frames. With $p = \text{COR}/100$, P_k is given by

$$P_k = \binom{N}{k} p^k (1-p)^{N-k}, \tag{9}$$

while R_k depends on the distribution of the estimation errors. The most favorable assumption would be to consider estimation errors uniformly distributed among all incorrect orientations. A more conservative assumption, which is supported by the experimental results, is that errors are distributed uniformly between 2 incorrect orientations. In this case, to $k < \lfloor \frac{N}{3} + 1 \rfloor$ we have $R_k = 0$; to $\lfloor \frac{N}{3} + 1 \rfloor \leq k < \lfloor \frac{N}{2} + 1 \rfloor$ we have $R_k = \sum_{m=N-2k+1}^{k-1} \binom{N-k}{m} 0.5^{N-k}$; and to $k \geq \lfloor \frac{N}{2} + 1 \rfloor$ we have $R_k = 1$.

Choosing $N = 70$ (which for a frame shift of 128 samples and 16 kHz sampling frequency, corresponds to a 560 ms delay) the values of COR_N that result from the values of COR columns in Table I are in Table II. As can be seen, there is a great improvement over single frame estimation.

VII. Conclusions

In this work, we studied the detection of the orientation of a loudspeaker, modeling a human speaker, using only two microphones. A large array could yield better results, but, inspired by the human auditory system, we restricted our analysis to two microphones. We showed that GMMs created with binaural cues and pairwise frame energies can be used to discriminate orientation. Knowing the user orientation can be beneficial when we have several voice-commanded devices. If we assume that the user turns to the equipment he wants to control, we can detect the orientation of the user and then select only the desired device. Finally, we extend the frame detection case to the multiframe case. Using multiframe estimation, a performance evaluation showed that a COR close to 100% is feasible for all orientations except for $180°$.

VIII. Acknowledgment

This work was partially supported by FAPESP under Grant 2010/18180-7.

References

[1] L. J. Griffiths and C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. on Antennas and Propagation*, 30(1):27–34, January 1982.

[2] J-M. Valin, F. Michaud, and J. Rouat. Robust 3d localization and tracking of sound sources using beamforming and particle filtering. *Proceedings of ICASSSP*, pages IV 841–844, 2006.

[3] S. Boll. Supression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.*, 27(2):113–120, April 1979.

[4] G. M. Davis. *Noise reduction in speech applications*. CRC press, Boca Raton, 2002.

[5] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitute estimator. *IEEE Trans. Acoust. Speech Signal Process.*, 32(6):1109–1121, December 1984.

[6] H. Lebart and J. M. Boucher. A new method based o spectral subtraction for speech enhancement. *Acta Acustica*, 83, 1997.

[7] E. Hansler and G. Schmidt. *Speech and Audio Processing in Adverse Environmentss*. Springer, Berlin, 2008.

[8] C. Avendano and H. Hermansky. Study on the dereverberation of speech based on temporal envelope filtering. *Proceedings of ICSLP*, 2:889–892, Oct. 1996.

[9] E. A. P. Habets. Multi-channel speech dereverberation based on a statistical model of late reverberation. *Proceedings of ICASSP*, pages 173–176, March 2005.

[10] C. Segura, C. C-Ferrer, A. Abad, J. R. Casas, and J. Hernando. Multimodal head orientation towards attention tracking in smartrooms. *Proceedings of ICASSP*, pages II 681–684, 2007.

[11] S Hwang, Y. Park, and Y. Park. Sound direction estimation using an artificial ear for robots. *Robotics and Autonomous Systems*, 59(3–4):208–217, March 2011.

TABLE I

PERFORMANCE OF THE GMM-BASED METHOD IN SINGLE FRAME CASE.

	ITD		E+ITD		ILD+ITD		E+ILD+ITD	
Angle(°)	COR(%)	AOE(°)	COR(%)	AOE(°)	COR(%)	AOE(°)	COR(%)	AOE(°)
0	88.0	16.7	98.3	2.5	95.3	6.6	95.5	6.4
45	84.9	14.8	91.4	8.1	90.7	8.9	91.4	8.2
90	83.9	9.7	80.1	13.6	79.1	14.4	80.4	13.7
135	66.0	19.9	60.2	23.6	63.7	22.4	58.1	24.9
180	29.5	54.1	48.3	41.7	43.6	43.3	48.1	41.4
225	65.5	29.4	58.1	38.1	63.4	31.7	66.6	29.3
270	86.9	7.3	88.4	7.5	88.7	7.0	88.8	6.8
315	97.2	2.2	97.8	1.2	96.7	2.1	96.9	1.9
Avg.	75.2	19.3	77.8	17.0	77.6	17.0	78.2	16.6

TABLE II

PERFORMANCE OF THE GMM-BASED METHOD IN MULTIFRAME CASE.

	ITD	E+ITD	ILD+ITD	E+ILD+ITD
Angle(°)	COR(%)	COR(%)	COR(%)	COR(%)
0	100	100	100	100
45	100	100	100	100
90	100	100	100	100
135	99.9	99.9	99.9	99.9
180	12.0	97.2	88.0	97.0
225	99.9	99.9	99.9	99.9
270	100	100	100	100
315	100	100	100	100
Avg.	89.0	99.6	98.5	99.6

[12] A. Abad, D. Macho, C. Segura, J. Hernando, and C. Nadeu. Effect of head orientation on the speaker localization performance in smart-room environment. *Proceedings of Interspeech*, pages 145–148, September 2005.

[13] A. Nakano, S. Nakagawa, and K. Yamamoto. Distant speech recognition using a microphone array network. *IEICE Transaction on Information and Systems*, E93-D(9):2451–2462, September 2010.

[14] A. Sasou. Acoustic head orientation estimation applied to powered wheelchair control. *2nd Int. ICST Conf. on Robot Communication and Coordination*, pages 1–6, May 2009.

[15] A. Y. Nakano, S. Nakagawa, and K. Yamamoto. Automatic estimation of position and orientation of an acoustic source by a microphone array network. *J. Acoust. Soc. Am.*, 126(6):3084–3094, December 2009.

[16] J. M. Sachar and H. F. Siverman. A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array. *Proceedings of ICASSSP*, pages IV 65–68, 2004.

[17] A. Brutti, M. Omologo, and P. Svaizer. Inference of acoustic source directivity using environment awareness. *Proceedings of 19th European Signal Processing Conference*, pages 151–155, 2011.

[18] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech, Signal Processing*, ASSP-24(4):320–327, August 1976.

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

[20] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. Speech and Audio Processing*, 3(1):72–83, 1995.