Robust Speaker's Location Detection in a Vehicle Environment Using GMM Models

Jwu-Sheng Hu, *Member, IEEE*, Chieh-Cheng Cheng, and Wei-Han Liu

***Abstract*—Human–computer interaction (HCI) using speech communication is becoming increasingly important, especially in driving where safety is the primary concern. Knowing the speaker's location (i.e., speaker localization) not only improves the enhancement results of a corrupted signal, but also provides assistance to speaker identification. Since conventional speech localization algorithms suffer from the uncertainties of environmental complexity and noise, as well as from the microphone mismatch problem, they are frequently not robust in practice. Without a high reliability, the acceptance of speech-based HCI would never be realized. This work presents a novel speaker's location detection method and demonstrates high accuracy within a vehicle cabinet using a single linear microphone array. The proposed approach utilize Gaussian mixture models (GMM) to model the distributions of the phase differences among the microphones caused by the complex characteristic of room acoustic and microphone mismatch. The model can be applied both in near-field and far-field situations in a noisy environment. The individual Gaussian component of a GMM represents some general location-dependent but content and speaker-independent phase difference distributions. Moreover, the scheme performs well not only in nonline-of-sight cases, but also when the speakers are aligned toward the microphone array but at difference distances from it. This strong performance can be achieved by exploiting the fact that the phase difference distributions at different locations are distinguishable in the environment of a car. The experimental results also show that the proposed method outperforms the conventional multiple signal classification method (MUSIC) technique at various SNRs.**

***Index Terms*—Gaussian mixture models (GMM), human–computer interaction (HCI), microphone array, sound localization.**

I. Introduction

ELECTRONIC devices for home, car, and personal applications are becoming more intelligent. One of the demands for intelligence is to enhance the convenience of operation, e.g., human–computer interaction (HCI) interfaces using speech communication [1]–[3]. Speech-based HCI is particularly important when the use of hands and eyes puts the user in danger. For example, the hands-free operation of in-car electronics is necessary during high-speed driving to avoid accidents. However, speech communication, unlike push-button operation, suffers from problems of unreliability because of environmental noise. Therefore, many speech enhancement techniques that use microphone arrays [4]–[7] have been introduced to enhance speech signals for a robust speech recognition system.

Speech communication involves the speakers' locations as well as speech recognition. Knowing the speaker's location is useful in determining who is talking (e.g., speaker identification). For example, in vehicle applications, a driver may wish to exert a particular authority in manipulating the in-car electronic systems. Additionally, for speech signal purification, a better receiving beam using a microphone array can be formed to suppress the environmental noises if the speaker's location is known. In a highly reflective or scattering environment, conventional delay estimation methods such as GCC-based algorithms [8]–[10] or previous works [11], [12] do not yield satisfactory results. Although Brandstein *et al.* [13] proposed Tukey's Biweight to redefine the weighting function to deal with the reflection effect; it is not suitable for a noisy environment. To overcome this limitation, Nikias *et al.* [14] adopted the alpha-stable distribution, instead of a single Gaussian model, to model ambient noise and to obtain a robust speaker's location detection in advance. In recent years, several works have introduced probability-based methods to eliminate the measurement errors caused by uncertainties, such as those associated with reverberation or low energy segments. Histogram-based time-delay of arrival (TDOA) estimators such as time histograms [15] and weighted time-frequency histograms [16], [17] have been proposed to reduce direction-of-arrival root-mean square errors. The algorithm in [17] performs well especially under low signal-to-noise ratio (SNR) conditions. Moreover, Potamitis *et al.* [18] proposed the probabilistic data association (PDA) technique with the interacting multiple model (IMM) estimator to conquer these measurement errors. Ward *et al.* [19] developed a particle filter beamforming in which the weights and particles can be updated using a likelihood function to solve the reverberation problem. These statistical-based methods [16]–[19] can improve the estimation accuracy further.

Another approach, proposed by Balan *et al.* [20], explores the eigenstructure of the correlation matrix of the microphone array by separating speech signals and noise signals into two orthogonal subspaces. The direction-of-arrival (DOA) is then estimated by projecting the manifold vectors onto the noise subspace. MUSIC [21], [22] combined with spatial smoothing [23] is one of the most popular methods for eliminating the coherence problem. However, as the experiment in this work indicates, its robustness is still poor in a real environment when the SNR is low. Furthermore, the near-field effect [24]–[26] should also be considered in applications in real environments.

Manuscript received January 27, 2005; revised June 22, 2005. This work was supported in part by the National Science Council of Taiwan, R.O.C., under Grant NSC 93-2218-E-009-031 and by the Ministry of Education, Taiwan, under Grant 91-1-FA06-4-4. This paper was recommended by Associate Editor S. Singh.

The authors are with the Department of Electrical and Control Engineering, National Chiao-Tung University, Hsinchu 300, Taiwan, R.O.C. (e-mail: jshu@nctu.edu.tw).

Digital Object Identifier 10.1109/TSMCB.2005.859084

In some environments, especially in vehicle environments, the line-of-sight condition may not be available because, for example, barriers may exist between the speaker and the microphone array. Consequently, when a single linear array is employed, the aforementioned methods cannot distinguish speakers under nonline-of-sight conditions. Hence, multiple microphone arrays must be considered [27], [28]. Further, the microphone mismatch problem often arises when such methods as GCC or eigenstructure-based algorithms are used since these methods require the microphones to be calibrated in advance. Accurate calibration is not easy to obtain since the characteristics of microphones vary from the sound source directions.

The relationship between a sound source and a receiver (microphone) in a complicated enclosure is almost impossible to characterize with a finite-length data in real-time applications (such as in frame-based calculations). According to the investigation of room acoustics [29], the number of eigen-frequencies with an upper limit of $f_s/2$ kHz can be obtained by the following equation:

$$L = \frac{4\pi}{3} B \left(\frac{f_s}{2\nu} \right)^3 \tag{1}$$

where f_s denotes the sampling frequency, ν represents the sound velocity ($\nu \approx 340$ m/s), and B is the geometrical volume. This equation indicates that the number of poles is too high when the frequency is high, and that the transient response occurs in almost any processing duration when the input signal is a speech signal. For example, the number of poles is about 96 435 when the sampling frequency is 8 kHz and the volume is 14.1385 m^3. Hence, the nonstationary characteristics of speech signals make the phase differences between the signals received by two elements of a microphone array from a fixed sound source vary among data sets. Moreover, the stochastic nature of the phase difference is more prominent when the source is moving slightly and environmental noise is present. Therefore, this work proposes the use of the distributions of phase differences, rather than their actual values, to locate the source, because the phase difference distributions vary among locations and can be distinguished by pattern matching methods. Previous research [30], [31] also showed that common acoustic measures vary significantly with small spatial displacements of the source or the microphone. The experimental results indicate that the Gaussian mixture model (GMM) [32] is very suitable for modeling these distributions. Furthermore, the model training uses the distributions of phase differences among microphones as a location-dependent but content and speaker-independent feature. In this case, the geometry of the microphone array should be considered to cope with the aliasing problem and maximize the phase difference of each frequency band to detect the speakers' locations accurately. Consequently, the microphone array can be decoupled into several pairs with various distances between the microphones to deal with different frequency bands. The location detector integrates the overall probability information from different frequency bands to detect the speakers' locations.

The environment of a vehicle raises all of the aforementioned issues of nonideality and speech-based HCI for in-car electronic systems such as mobile phones, navigation devices and stereos is necessary to enhance driving safety. Therefore, the experimental verification of this paper was performed in a vehicle. The remainder of this work is organized as follows. The following section discusses the system architecture and the relationship between the selected frequency and geometry of the microphone array. Section III presents the procedure for training of the Gaussian mixture location model using the EM algorithm and the location detection method. Section IV experimentally compares the proposed approach to the conventional MUSIC method. Conclusions are drawn in Section V.

II. System Architecture and Microphone Array Signal Processing

A. System Architecture

Fig. 1 illustrates the overall system architecture. A voice activity detector separates the system into two stages, the silent stage and the speech stage, where the voice activity detection algorithms could be found in [33], [34]. The first stage is called the silent stage in which speakers are silent. In this stage, online environmental noise without speech was recorded. Noise is assumed to be additive, so the signal received when a speaker is talking in a car can be expressed as a linear combination of the speech signal and the environmental noise. Therefore, in this stage, the system combines the online recorded environmental noise, $N_1(\omega), \cdots, N_M(\omega)$, with the pre-recorded speech database, $S_1(\omega), \cdots, S_M(\omega)$, to construct training signals, $X_1(\omega), \cdots, X_M(\omega)$, and to derive the GM location models, where M denotes the number of microphones. Consequently, a set of pre-recorded speech data at different location was required to obtain a priori information between the speaker and the microphone array. The pre-recorded speech database was collected at each location when the environment was quiet and can represent the acoustical characteristic of each location. The distributions of phase differences, which include information on mismatches among microphones, can be modeled based on the *a priori* information. Since the environmental noise alters, the GM location models that contain the characteristics of environmental noise are updated in this stage for robustness. The second stage is the speech stage, in which the parameters of GM location models derived from the first stage are duplicated into the location detector to detect the speaker's location.

B. Frequency Band Divisions Based on a Uniform Microphone Array

The phase difference of the received signal becomes more significant as the distance between microphones increases. However, the aliasing problem occurs when this distance exceeds half of the maximum wavelength of the received signal [35]. The distance between pairs of microphones should be chosen based on the selected frequency band to obtain clear phase difference data to enhance the accuracy of location detection and prevent aliasing. Accordingly, Fig. 2 illustrates a uniform microphone array with M microphones and the distance of d.

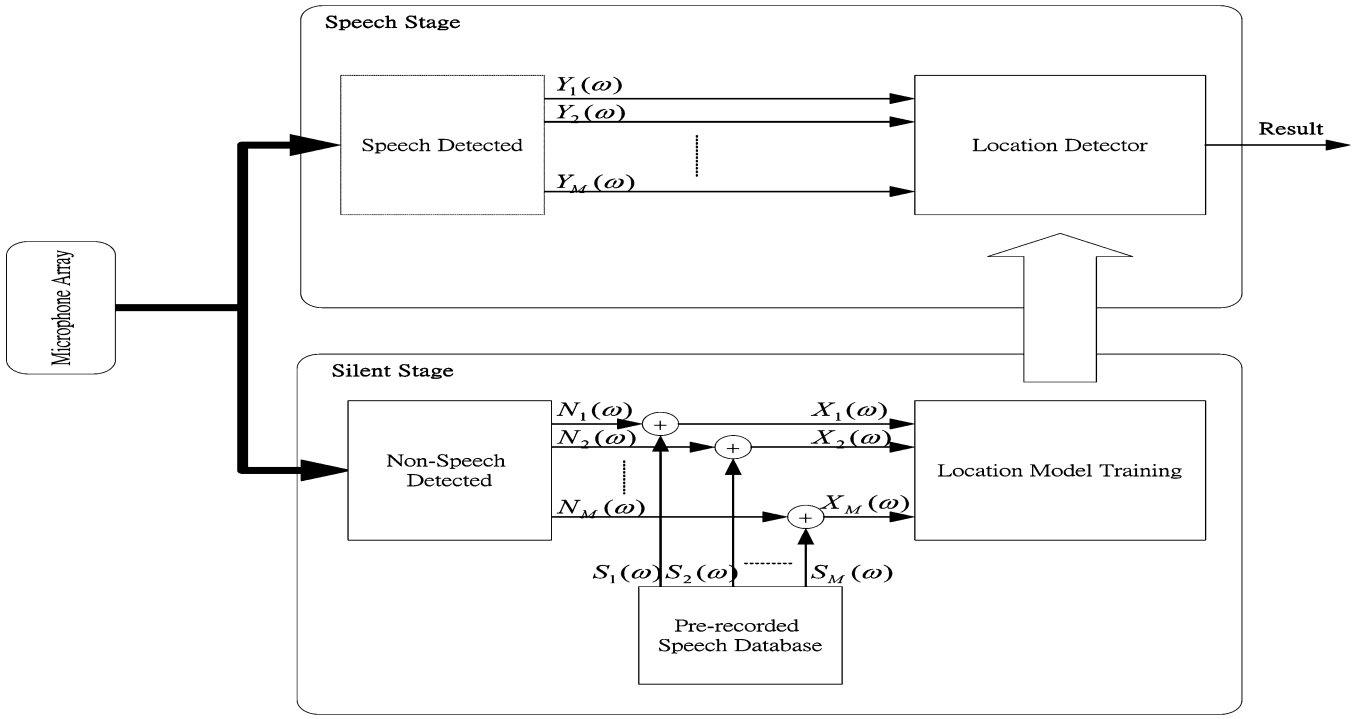


Fig. 1. Overall system architecture.

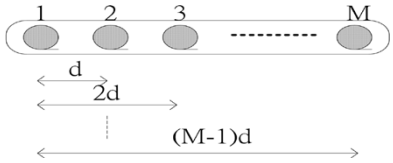


Fig. 2. Microphone array geometry.

Based on this geometry, the processed frequencies can be divided into at most $(M-1)$ bands. Table I lists the different frequency bands and the corresponding microphone pairs, where m denotes the mth microphone; b represents the band number and J_b is the number of microphone pairs in the band b. Each frequency component belonging to a specific frequency band of any location joints the phase differences measured by the microphone pairs to generate a self-GMM.

III. The Gaussian Mixture Location Model

A. Location Model Description

A Gaussian mixture density in the band b is a weighted sum of N Gaussian component densities, and is denoted as

$$G_b\left(\boldsymbol{P}_x(\omega,b)|\boldsymbol{\lambda}(\omega,b)\right)=\sum_{i=1}^{N}\rho_i(\omega,b)g_i\left(\boldsymbol{P}_x(\omega,b)\right) \quad (2)$$

where $\boldsymbol{P}_x(\omega,b)=[P_x(\omega,1)\ \cdots\ P_x(\omega,J_b)]^T$ is a J_b -dimensional phase difference vector combined with pre-recorded database and environmental noise. The phase difference can be obtained as follows:

$$P_x(\omega,j)=\text{phase}\left(X_{j+M-b}(\omega)\right)-\text{phase}\left(X_j(\omega)\right) \quad \text{with } 1\le j\le b \quad (3)$$

where $\rho_i(\omega,b)$ is the ith mixture weight, and $\boldsymbol{\lambda}(\omega,b)$ represents the complete Gaussian mixture density which is parameterized by the mean vector, covariance matrices and mixture weights from N component densities

$$\boldsymbol{\lambda}(\omega,b)=\{\boldsymbol{\rho}(\omega,b),\boldsymbol{\mu}(\omega,b),\boldsymbol{\Sigma}(\omega,b)\} \quad (4)$$

where $\boldsymbol{\rho}(\omega,b)=[\rho_1(\omega,b)\ \cdots\ \rho_N(\omega,b)]$ denotes the mixture weight vector in the band b, $\boldsymbol{\mu}(\omega,b)=[\boldsymbol{\mu}_1(\omega,b)\ \cdots\ \boldsymbol{\mu}_N(\omega,b)]$ denotes the mean matrix in the band b, and $\boldsymbol{\Sigma}(\omega,b)=[\boldsymbol{\Sigma}_1(\omega,b)\ \cdots\ \boldsymbol{\Sigma}_N(\omega,b)]$ denotes the covariance matrix in the band b. The corresponding vector and matrix of the parameters defined above are

$$\boldsymbol{\mu}_i(\omega,b)=[\mu_i(\omega,1)\quad\cdots\quad\mu_i(\omega,J_b)]^T$$

$$\boldsymbol{\Sigma}_i(\omega,b)=\begin{bmatrix}\sigma_i^2(\omega,1) & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \sigma_i^2(\omega,J_b)\end{bmatrix}.$$

$g_i(\boldsymbol{P}_x(\omega,b))$ denotes the Gaussian component density

$$g_i\left(\boldsymbol{P}_x(\omega,b)\right)=\frac{1}{(2\pi)^{\frac{J_b}{2}}\left|\boldsymbol{\Sigma}_i(\omega,b)\right|^2}*\exp\left(-\frac{1}{2}\left[\boldsymbol{P}_x(\omega,b)-\boldsymbol{\mu}_i(\omega,b)\right]^T\boldsymbol{\Sigma}_i(\omega,b)^{-1}\left[\boldsymbol{P}_x(\omega,b)-\boldsymbol{\mu}_i(\omega,b)\right]\right). \quad (5)$$

The mixture weight must satisfy the constraint that

$$\sum_{i=1}^{N}\rho_i(\omega,b)=1. \quad (6)$$

Each band associated with each location is represented by a GMM and is referred to by its parameter matrix, $\boldsymbol{\lambda}(\omega,b)$. The covariance matrix, $\boldsymbol{\Sigma}_i(\omega,b)$, is selected as a diagonal matrix. Even though the phase differences of the microphone pairs may

TABLE I
FREQUENCY BANDS CORRESPOND TO THE MICROPHONE PAIRS

Frequency band	Microphone pairs	The number of microphone pair	The range of frequency band
Band 1 ($b=1$)	$(m, m+M-1)$ with $m=1$	$J_1=1$	$0 \le f \le \frac{v}{2(M-1)d}$
Band 2 ($b=2$)	$(m, m+M-2)$ with $1 \le m \le 2$	$J_2=2$	$\frac{v}{2(M-1)d} < f \le \frac{v}{2(M-2)d}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Band $M-1$ ($b=M-1$)	$(m, m+1)$ with $1 \le m \le M-1$	$J_{M-1}=M-1$	$\frac{v}{4d} < f \le \frac{v}{2d}$

not be statistically independent of each other, GM models with diagonal covariance matrices have been observed to be capable of modeling the correlations within the data by utilizing larger mixture numbers [36].

B. Parameters Estimation via EM Algorithm

The purpose is to determine the parameters of the GMM, $\boldsymbol{\lambda}(\omega, b)$, from the measured phase differences between each microphone pair in band b. Several techniques are available for estimating $\boldsymbol{\lambda}(\omega, b)$, of which the most popular is the EM algorithm [32] that estimates the parameters by using an iterative scheme to maximum the log-likelihood function. The EM algorithm can guarantee a monotonic increase in the model's log-likelihood value, and its iteration equations corresponding to frequency band selection can be arranged as follows.

Expectation step:

$$G_b\left(i|\boldsymbol{P}_x^{(t)}(\omega,b),\boldsymbol{\lambda}(\omega,b)\right)=\frac{\rho_i(\omega,b)g_i\left(\boldsymbol{P}_x^{(t)}(\omega,b)\right)}{\sum_{i=1}^{N}\rho_i(\omega,b)g_i\left(\boldsymbol{P}_x^{(t)}(\omega,b)\right)} \quad (7)$$

where $G_b(i|\boldsymbol{P}_x^{(t)}(\omega,b),\boldsymbol{\lambda}(\omega,b))$ is an *a posteriori* probability and $\mathbf{P}_x(\omega,b) = \{\boldsymbol{P}_x^{(1)}(\omega,b),\cdots,\boldsymbol{P}_x^{(T)}(\omega,b)\}$ is a sequence of T input phase difference vectors.

Maximization step:

i) Estimate the mixture weights

$$\rho_i(\omega,b)=\frac{1}{T}\sum_{t=1}^{T}G_b\left(i|\boldsymbol{P}_x^{(t)}(\omega,b),\boldsymbol{\lambda}(\omega,b)\right). \quad (8)$$

ii) Estimate the mean vector

$$\boldsymbol{\mu}_i(\omega,b)=\frac{\sum_{t=1}^{T}G_b\left(i|\boldsymbol{P}_x^{(t)}(\omega,b),\boldsymbol{\lambda}(\omega,b)\right)\boldsymbol{P}_x^{(t)}(\omega,b)}{\sum_{t=1}^{T}G_b\left(i|\boldsymbol{P}_x^{(t)}(\omega,b),\boldsymbol{\lambda}(\omega,b)\right)}. \quad (9)$$

iii) Estimate the variances

$$\sigma_i^2(\omega,j)=\frac{\sum_{t=1}^{T}G_b\left(i|\boldsymbol{P}_x^{(t)}(\omega,b),\boldsymbol{\lambda}(\omega,b)\right)\boldsymbol{P}_x^{(t)^2}(\omega,j)}{\sum_{t=1}^{T}G_b\left(i|\boldsymbol{P}_x^{(t)}(\omega,b),\boldsymbol{\lambda}(\omega,b)\right)}-\mu_i^2(\omega,j) \quad (10)$$

where $j=\{1,\cdots,J_b\}$.

However, the EM algorithm only guarantees to find a local maximum log-likelihood model. A different choice of initial model $\boldsymbol{\lambda}_0(\omega,b)$ leads to various local maximum models. This work considered two initialization methods to find out the initial model. K-means [37] is by far the most widely used method. Charles [38] proposed an accelerated K-means algorithm which utilizes the triangle inequality to decrease significantly the computational power requirement. Charles' method is also suitable for finding a good initial model to lower the iteration number of the EM algorithm. The first method utilizes the accelerated K-means clustering method. The second method separates phase difference range $\{-\pi,\pi\}$ into N segments to obtain a fixed initial mean model since the phase difference range is small enough. Consequently, the initial mean model is $\{-\pi\ (2\pi/N-1)-\pi\ (4\pi/N-1)-\pi\ \cdots\ \pi\}$. The location detection performances of the two initial approaches have slightly different performance and no one is always the best. Fig. 3 shows the location model training procedure with the total location number L.

C. Location Detection

Assume a group of L locations, $\{1,\cdots,L\}$, which are represented by the parameters $\boldsymbol{\lambda}_1(\omega),\cdots\{\boldsymbol{\lambda}_L(\omega)$ with $\boldsymbol{\lambda}_l(\omega) = \{\boldsymbol{\lambda}_l(\omega,1),\cdots,\boldsymbol{\lambda}_l(\omega,M-1)\}$. The location is determined by finding the GM location model which has the maximum *posteriori* probability for a given observation sequences

$$\begin{aligned}\hat{l}&=\arg\max_{1\le l\le L}\sum_{b=1}^{M-1}\log\left[G_b\left(\boldsymbol{\lambda}_l(\omega,b)|\mathbf{P}_Y(\omega,b)\right)\right]\\&=\arg\max_{1\le l\le L}\sum_{b=1}^{M-1}\log\frac{G_b\left(\mathbf{P}_Y(\omega,b)|\boldsymbol{\lambda}_l(\omega,b)\right)p\left(\boldsymbol{\lambda}_l(\omega,b)\right)}{p\left(\mathbf{P}_Y(\omega,b)\right)}\end{aligned} \quad (11)$$

where $\mathbf{P}_Y(\omega,b)=\{\boldsymbol{P}_Y^{(1)}(\omega,b),\cdots,\boldsymbol{P}_Y^{(Q)}(\omega,b)\}$ is a phase difference testing sequence derived from $Y_1(\omega),\cdots,Y_M(\omega)$, and Q denotes the length of the testing sequence. If the probability densities at all locations are equally likely, then $p(\boldsymbol{\lambda}_l(\omega))$ could be chosen as $1/L$. The probability $p(\mathbf{P}_Y(\omega,b))$ is the same for all location models and the detection rule can be rewritten as

$$\hat{l}=\arg\max_{1\le l\le L}\sum_{b=1}^{M-1}\sum_{q=1}^{Q}\log G_b\left(\boldsymbol{P}_Y^{(q)}(\omega,b)|\boldsymbol{\lambda}_l(\omega,b)\right). \quad (12)$$

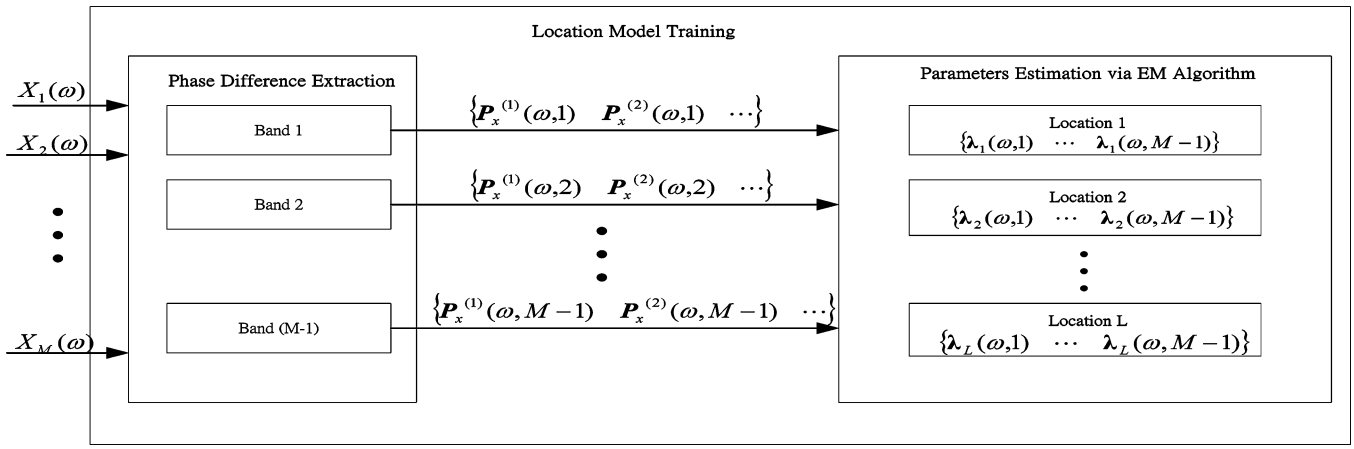


Fig. 3. Location model training procedure.

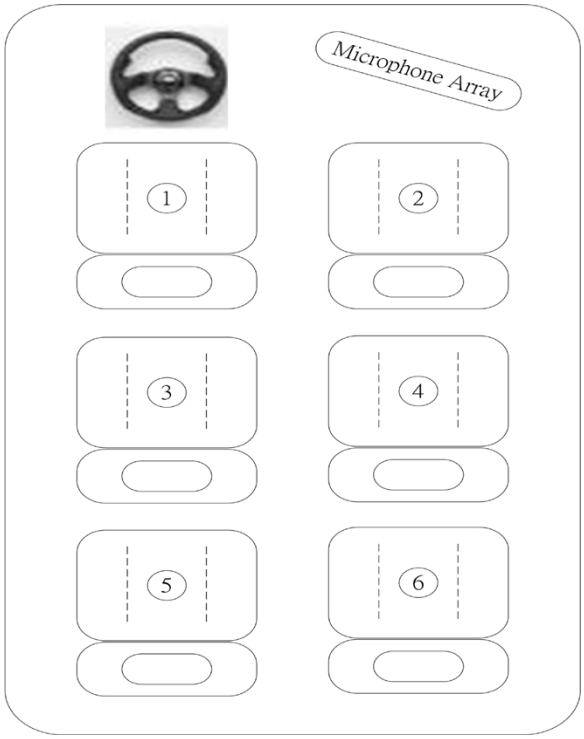


Fig. 4. Position number and microphone array position.

Fig. 5. Physical configuration in the vehicle.

TABLE II
SNR RANGES AT VARIOUS SPEEDS

Speed	Speed = 0 *Km/hr*	Speed = 20 *Km/hr*
The SNR range (dB)	10.8204 ~ 17.2664	4.1762 ~ 10.6222
Speed	Speed = 40 *Km/hr*	Speed = 60 *Km/hr*
The SNR range (dB)	-4.5320 ~ 1.9140	-6.2526 ~ 0.1934
Speed	Speed = 80 *Km/hr*	Speed = 100 *Km/hr*
The SNR range (dB)	-8.4709 ~ -2.0249	-13.0531 ~ -6.6071

IV. EXPERIMENTAL RESULTS

The experiment was performed in a mini-van vehicle with six separated seats [39]. Fig. 4 presents the locations of the seats. During the experiment, the speakers at these locations moved their bodies or heads to mimic real usage scenarios. A uniform linear array of six microphones with 5-cm spacing was mounted in front of location no. 2. Fig. 5 displays the physical configuration inside the vehicle. The environmental noise signals were varied as the car drove at various speeds. During the experiment, all windows were closed to prevent the microphones from saturating and the cabinet temperature was set to be 24 °C using the in-car air conditioner. Off-the-shelf, low-cost, and noncalibrated microphones were used for the array. The amplified microphone signals were digitized by 16-bit AD converters. Table II lists the SNR ranges at various speeds, corresponding to the six locations. Table III presents the frequency bands that correspond to the pairs of microphones. The voice activity detection algorithm provided in [33] was utilized in this experiment. The received signals were sampled at 8 kHz, and the window for the short-time Fourier transform (STFT) contained 256 zero padding samples and 32 ms speech signals, totaling 512 samples. Fig. 6 illustrates the processed frame and the overlapping condition. The experiment was performed in both quiet and noisy environments. Fig. 7 shows the testing signal in driver's seat in a quiet environment, and Fig. 8 depicts the testing signals at two different speeds, 40 and 100 km/h.

A. MUSIC Algorithm

A wideband incoherent MUSIC algorithm [22] with arithmetic mean was implemented and the results were compared with those of the proposed approach. Ten major frequencies, ranging from 0.1 to 3.4 kHz, were adopted for the MUSIC algorithm. Outliers were removed from the estimated angles by utilizing the method provided in [40]. Moreover, the angle errors

TABLE III
FREQUENCY BANDS CORRESPOND TO THE MICROPHONE PAIRS

Frequency band	Microphone pairs	The number of microphone pair	The range of frequency band
Band 1 ($b=1$)	(1,6)	$J_1=1$	$0 \le f \le 680Hz$
Band 2 ($b=2$)	(1,5); (2,6)	$J_2=2$	$680Hz < f \le 850Hz$
Band 3 ($b=3$)	(1,4); (2,5); (3,6)	$J_3=3$	$850Hz < f \le 1100Hz$
Band 4 ($b=4$)	(1,3); (2,4); (3,5); (4,6)	$J_4=4$	$1100Hz < f \le 1700Hz$
Band 5 ($b=5$)	(1,2); (2,3); (3,4); (4,5); (5,6)	$J_5=5$	$1700Hz < f \le 3400Hz$

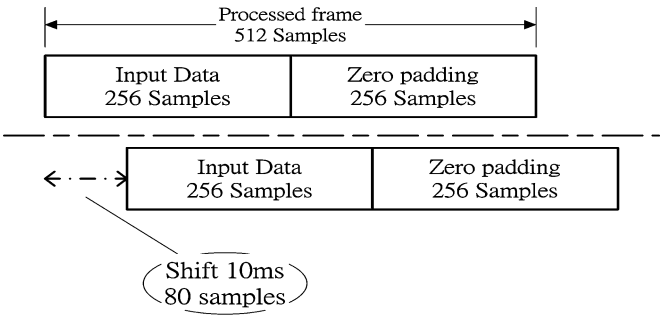


Fig. 6. Processed frame and overlapping condition.

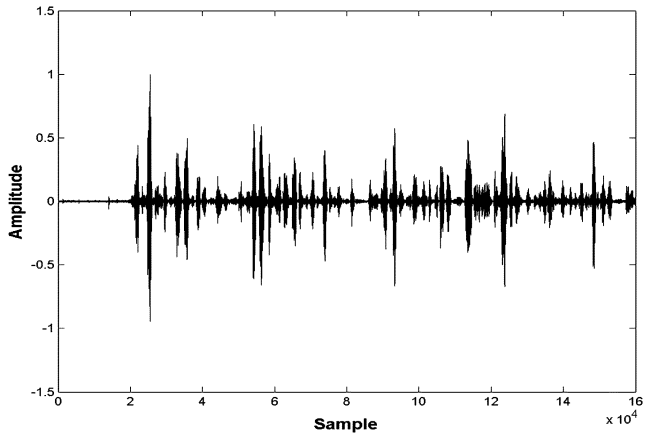


Fig. 7. Testing signal received at the driver's seat in a quiet environment.

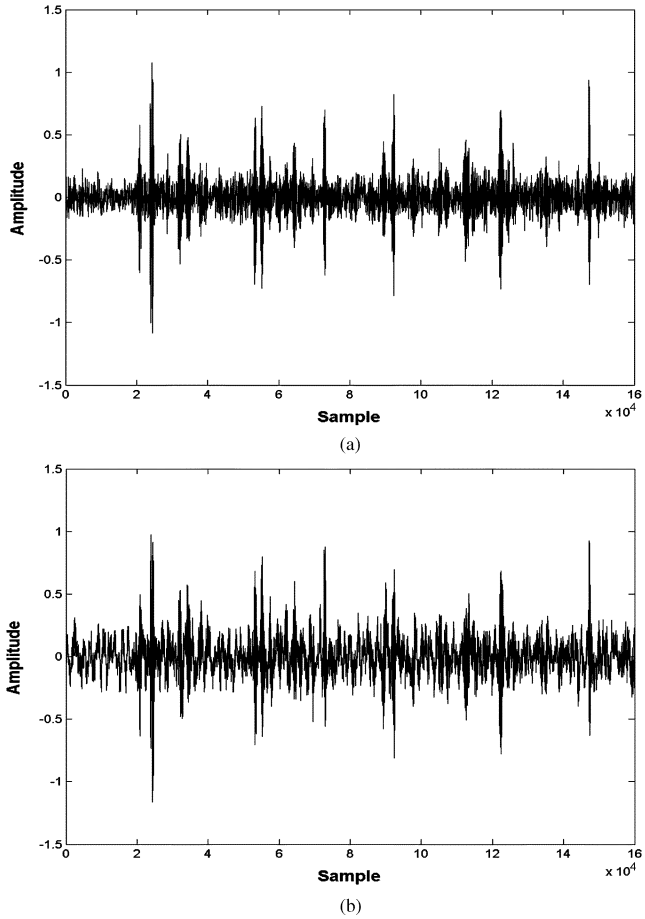


Fig. 8. Testing signal received at driver's seat in speed 40 and 100 km/h. (a) Speed is 40 km/h. (b) Speed is 100 km/h.

needed for outlier rejection was derived from the estimated angles and real angles. Table IV lists the real angles and the statistical results after the outlier have been rejected. It demonstrates that even in a quiet environment, the MUSIC method might not be sufficiently accurate. The standard deviations in Table IV at various speeds were too large to distinguish among them, especially at locations no. 3 and 5. In particular, the standard deviations at locations no. 4 and 6 exceeded those at other locations at almost all noise levels. This phenomenon could be caused by nonline-of-sight effect.

Locations no. 2, 4, and 6 had the same DOA to the microphone array. Therefore, only locations no. 1, 2, 3, and 5 were considered for online testing. A frequently used classification method, K-nearest-neighbor classification rule (KNN [41]) was used to construct a flexible boundary to improve the accuracy of detection to cope with the slight movement of the source, microphone mismatch, transient response, and environmental noise. The estimated angles following outlier rejection were used as reference data in online location detection to illustrate further the performance of the MUSIC algorithm in the car cabinet.

Suppose that the lth location contains β_l estimated angles and that $\sum_l \beta_l = \beta$ is the reference data set. Assume that the lth location contains K_l points in the K-*nearest* results of a new estimate $\hat{r}$ derived from MUSIC with outlier rejection. The *a posteriori* probability is then given as

$$p(l|\hat{r}) = \frac{K_l}{K}. \tag{13}$$

To minimize the probability of a false classification of $\hat{r}$, the estimated location, denoted as $\hat{l}_{\mathrm{MUSIC}}$, was decided by using the following equation:

$$\hat{l}_{\mathrm{MUSIC}} = \arg \max_{l=1,2,3,5} p(l|\hat{r}). \tag{14}$$

Notably, the new estimate was not classified if it is an outlier according to the results in Table IV. The parameters were set

TABLE IV
MEAN, STANDARD DEVIATION, AND AVERAGE OUTLIER PROBABILITY OF ESTIMATED ANGLES

		Location No.1	Location No.2	Location No.3	Location No.4	Location No.5	Location No.6	Average outlier probability
	Mean	-27.31	23.31	-8.21	19.44	-0.82	18.57	
	Real angle	-30	20	-14	20	0	20	
Standard deviations at various speeds	In a quiet environment	7.89	6.33	7.53	10.82	4.91	12.70	24.08
	Speed 0 *Km/hr*	8.17	7.19	6.43	12.30	4.58	11.34	28.55
	Speed 20 *Km/hr*	8.51	6.48	7.37	12.65	4.52	9.01	35.14
	Speed 40 *Km/hr*	10.96	6.72	6.42	11.40	4.95	9.52	45.72
	Speed 60 *Km/hr*	10.65	7.21	6.17	11.46	3.82	10.12	50.63
	Speed 80 *Km/hr*	9.68	7.91	6.13	11.13	3.85	8.69	54.16
	Speed 100 *Km/hr*	7.88	7.67	6.94	8.73	5.20	9.31	45.91

TABLE V
CORRECT RATE OF MUSIC METHOD UTILIZING KNN WITH OUTLIER REJECTION

Location	The correct rates at various speeds (*Km/hr*)						
	In a quiet environment	Speed 0 *Km/hr*	Speed 20 *Km/hr*	Speed 40 *Km/hr*	Speed 60 *Km/hr*	Speed 80 *Km/hr*	Speed 100 *Km/hr*
1	97 %	94 %	85 %	74 %	79 %	84 %	91 %
2	94 %	93 %	90 %	92 %	89 %	81 %	89 %
3	58 %	60 %	44 %	63 %	70 %	36 %	52 %
4	×	×	×	×	×	×	×
5	64 %	59 %	46 %	17 %	26 %	78 %	22 %
6	×	×	×	×	×	×	×

to $\beta_l = 200$, $l = \{1,2,3,5\}$, $\beta = 800$, and $K = 30$ and the number of trials was 100. Table V presents the correct rate after KNN classification. The correct rates at locations no. 3 and 5 were too low to be useful. The reason was the estimated means of angles were so close and the standard deviations were so large at locations no. 3 and 5 that they cannot be distinguished to each other (see Table IV). In summary, these experimental results demonstrate that the MUSIC algorithm is not sufficiently reliable in a vehicle environment, even a classification method is applied and outliers are rejected to cope with the uncertainties.

B. The Proposed Method

The proposed method was applied under the same experimental conditions as part A to detect the speaker's location. The second initial approach mentioned in Section III was utilized to initialize the mean values. The covariance update may lead to numerical difficulties, as the covariance matrices become nearly singular. Consequently, the practical solution is to limit the minimum variance $\sigma^2_{\min}$. In this experiment, the value of $\sigma^2_{\min}$ was set to 0.02. The lengths of the training sequence T and the testing sequence Q were set to 200 and 50; in other words, a two-second length input datum was set for training, and a half-second length input datum was set for testing. The mixture number of GMM model has ten choices, from one to ten. Fig. 9 plots the experimental result of the correct rate versus the mixture numbers at 100 km/h. As shown in Fig. 9(a), a single Gaussian distribution (where the mixture number is one) could not yield a satisfactory experimental performance. The correct rates were 100% at all locations when the mixture was ten. This finding justifies the assumption that GMM is suitable for this application. Although the experimental performance improved

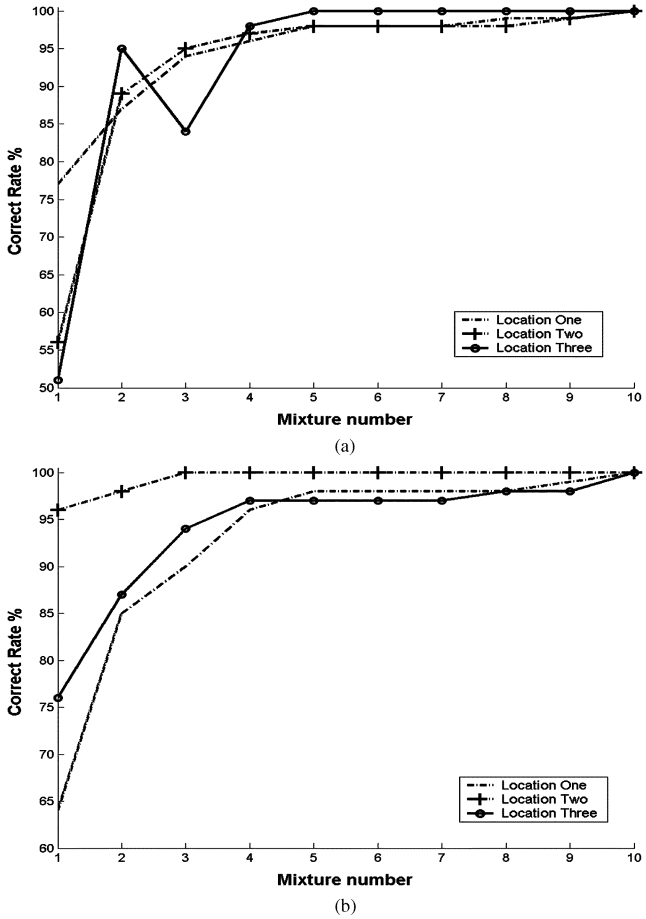


Fig. 9. Correct rate versus the different mixture numbers in 100 km/h. (a) Location number is chosen from 1 to 3. (b) Location number is chosen from 4 to 6.

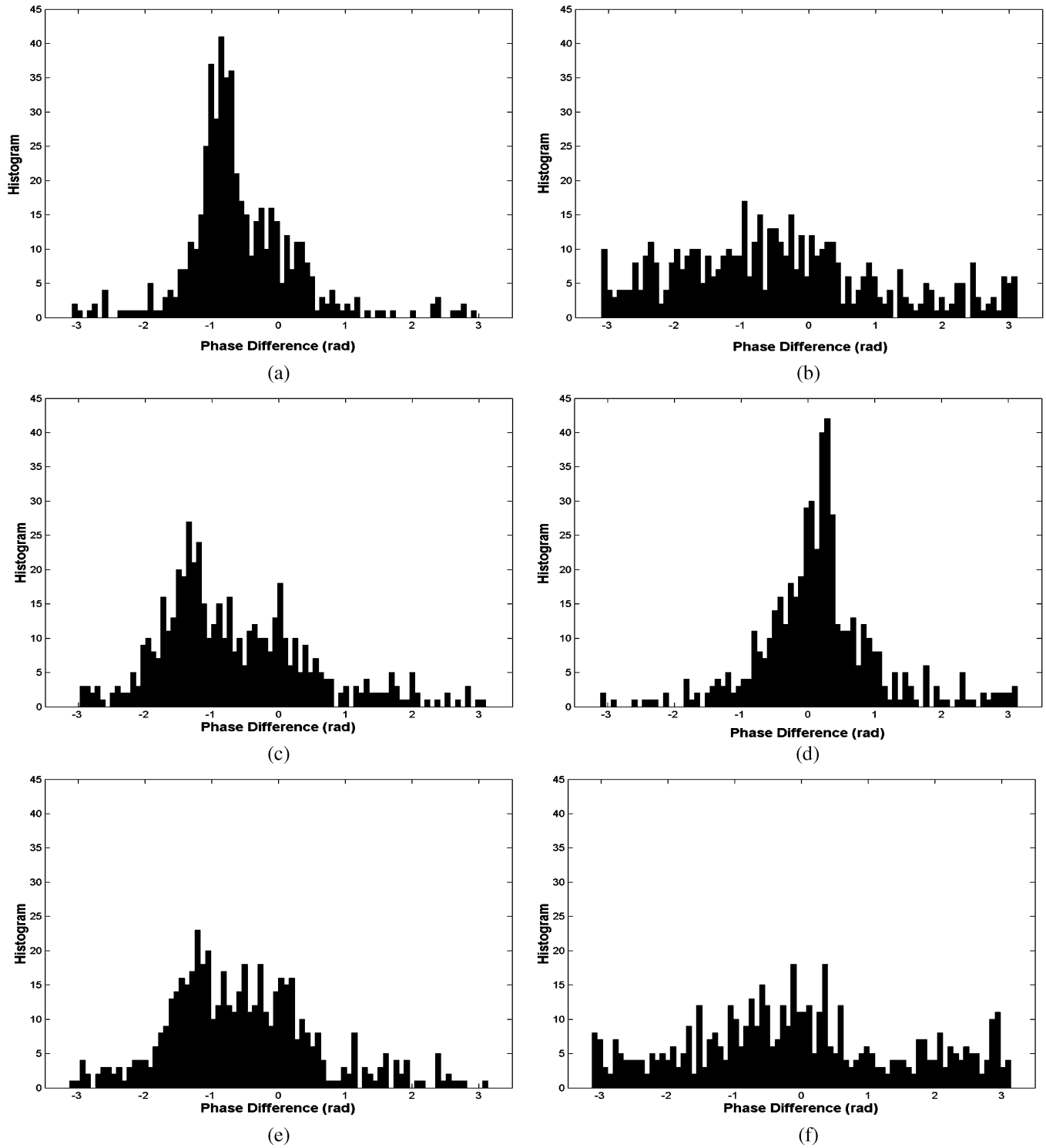


Fig. 10. The histograms of phase differences at locations no. 2, 4, and 6 between the third and the sixth microphones at a frequency of 0.9375 kHz and between the fourth and the sixth microphones at a frequency of 1.5 kHz, e.g., in the third and fourth frequency bands ($\mathrm{speed} = 100\ \mathrm{km/h}$). (a) Histogram of phase difference at location no. 2 ($\mathrm{frequency} = 0.9375\ \mathrm{kHz}$). (b). Histogram of phase difference at location no. 4 ($\mathrm{frequency} = 0.9375\ \mathrm{kHz}$). (c) Histogram of phase difference at location no. 6 ($\mathrm{frequency} = 0.9375\ \mathrm{kHz}$). (d) Histogram of phase difference at location no. 2 ($\mathrm{frequency} = 1.5\ \mathrm{kHz}$). (e) Histogram of phase difference at location no. 4 ($\mathrm{frequency} = 1.5\ \mathrm{kHz}$). (f) Histogram of phase difference at location no. 6 ($\mathrm{frequency} = 1.5\ \mathrm{kHz}$).

as the mixture number increased, the improvement in performance was not significant when the mixture number exceeded five. Table VI lists the experimental results with a mixture number of five. Clearly, the proposed method outperforms the MUSIC algorithm. Even at locations no. 4 and 6, the proposed method could distinguish them with significant accuracy. Fig. 10 shows the histograms of phase differences at locations no. 2, 4, and 6 between the third and the sixth microphones at a frequency of 0.9375 kHz and between the fourth and the sixth microphones at a frequency of 1.5 kHz, e.g., in the third and fourth frequency bands. The speed that corresponds to this figure is 100 km/h. Although the locations had the same angle

TABLE VI
EXPERIMENTAL RESULT OF THE PROPOSED METHOD WITH A MIXTURE NUMBER OF FIVE

Location	The correct rates at various speeds (*Km/hr*)						
	In a quiet environment	Speed 0 *Km/hr*	Speed 20 *Km/hr*	Speed 40 *Km/hr*	Speed 60 *Km/hr*	Speed 80 *Km/hr*	Speed 100 *Km/hr*
1	100 %	99 %	100 %	99 %	99 %	99 %	98 %
2	100 %	99 %	100 %	99 %	99 %	97 %	98 %
3	100 %	100 %	100 %	99 %	100 %	100 %	100 %
4	100 %	99 %	99 %	99 %	98 %	98 %	98 %
5	100 %	100 %	100 %	100 %	100 %	100 %	100 %
6	100 %	99 %	99 %	98 %	99 %	98 %	97 %

to the microphone array, their phase difference distributions were quite different, as indicated by several research reports [30], [31]. Additionally, the proposed method combined five frequency bands, each of which contained different phase difference distributions. As a result, the proposed method was able to distinguish all of the locations by exploiting their implicit diversities. Moreover, under low SNR conditions, the proposed approach still yielded a high correct rate and was robust against in-vehicle noise.

V. CONCLUSION

This paper proposes a robust speaker location detection method. The proposed method can overcome practical issues, such as the microphone mismatch, near-field effect, local scattering, and coherence problems. Additionally, the proposed method was found out to work even under nonline-of-sight conditions and when speakers are in the same direction but different distances from the microphone array. This investigation also presents a systematic procedure to improve the ability of environmental adaptation. The experimental results show the robustness and accuracy of the proposed method even under a low SNR. The proposed localization method has potential to be applied to other wave media, as their behaviors are similar to that of an acoustic wave in an enclosure.

REFERENCES

[1] K. Pulasinghe, K. Watanabe, K. Izumi, and K. Kiguchi, "Modular fuzzy-neuro controller driven by spoken language commands," *IEEE Trans. Syst., Man, Cybern. B*, vol. 34, no. 1, pp. 293–302, Feb. 2004.

[2] J. A. Borges, J. Jimenez, and N. J. Rodriquez, "Speech browsing the World Wide Web," *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*, vol. 4, pp. 12–15, Oct. 1999.

[3] R. Sharma, V. I. Pavlovic, and T. S. Huang, "Toward multimodal human-computer interface," *Proc. IEEE*, vol. 86, no. 5, pp. 853–869, May 1998.

[4] P. Aarabi and S. Guangji, "Phase-based dual-microphone robust speech enhancement," *IEEE Trans. Syst., Man, Cybern. B*, vol. 34, no. 4, pp. 1763–1773, Aug. 2004.

[5] M. Dahl and I. Claesson, "Acoustic noise and echo cancelling with microphone array," *IEEE Trans. Veh. Technol.*, vol. 48, no. 5, pp. 1518–1526, Sep. 1999.

[6] J. G. Ryan and R. A. Goubran, "Application of near-field optimum microphone arrays to hands-free mobile telephony," *IEEE Trans. Veh. Technol.*, vol. 52, no. 2, pp. 390–400, Mar. 2003.

[7] J. Hu, C. C. Cheng, and W. H. Liu, "Processing of speech signals using microphone array for intelligent robots," *Proc. Inst. Mech. Eng.I: J. Syst. Contr. Eng.*, vol. 219, pp. 133–144, 2005.

[8] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The Smoothed Coherence Transform(SCOT)," Naval Underwater Systems Center, New London Lab., New London, CT, Tech. Memo TC-159-72, Aug. 8, 1972.

[9] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 8, pp. 320–327, Aug. 1976.

[10] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform," *IEEE Signal Process. Lett.*, vol. 61, pp. 1497–1498, Oct. 1973.

[11] J. Hu, T. M. Su, C. C. Cheng, W. H. Liu, and T. I. Wu, "A self-calibrated speaker tracking system using both audio and video data," in *Proc. IEEE Conf. Control Applications*, vol. 2, Sep. 2002, pp. 731–735.

[12] J. Hu, C. C. Cheng, W. H. Liu, and T. M. Su, "A speaker tracking system with distance estimation using microphone array," in *Proc. IEEE/ASME Int. Conf. Advanced Manufacturing Technologies and Education*, Aug. 2002.

[13] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, Apr. 1997, pp. 375–378.

[14] C. L. Nikas and M. Shao, *Signal Processing with Alpha-Stable Distributions and Applications*. New York: Wiley, 1995.

[15] N. Strobel and R. Rabenstein, "Classification of time delay estimates for robust speaker localization," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 6, Mar. 1999, pp. 15–19.

[16] S. Mavandadi and P. Aarabi, "Multichannel nonlinear phase analysis for time-frequency data fusion," in *Proc. SPIE, Architectures, Algorithms, and Applications VII (AeroSense 2003)*, vol. 5099, Apr. 2003, pp. 222–231.

[17] P. Aarabi and S. Mavandadi, "Robust sound localization using conditional time-frequency histograms," *Inform. Fusion*, vol. 4, pp. 111–122, Jun. 2003.

[18] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 520–529, Sep. 2004.

[19] D. B. Ward and R. C. Williamson, "Particle filter beamforming for acoustic source localization in a reverberant environment," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, May 2002, pp. 1777–1780.

[20] R. V. Balan and J. Rosca, "Apparatus and method for estimating the direction of arrival of a source signal using a microphone array," European Patent US2 004 013 275, 2004.

[21] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.

[22] M. Wax, T. Shan, and T. Kailath, "Spatio-Temporal spectral analysis by eigenstructure methods," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 4, pp. 817–827, Aug. 1984.

[23] H. Wang and M. Kaveh, "Coherent signal-subspace processing for detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 4, pp. 823–831, Aug. 1985.

[24] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *IEEE Trans. Syst., Man, Cybern. B*, vol. 34, pp. 1526–1540, Jun. 2004.

[25] J. G. Ryan and R. A. Goubran, "Array optimization applied in the near field of a microphone array," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 173–176, Mar. 2000.

[26] Y. R. Zheng, R. A. Goubran, and M. K. El-Tanany, "Robust near-field adaptive beamforming with distance discrimination," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 12, no. 9, pp. 478–488, Sep. 2004.

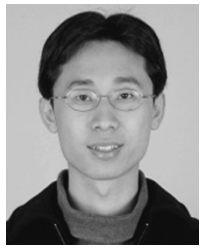
[27] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 45–50, Sep. 1997.

[28] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 288–292, May 1997.

[29] H. Kuttruf, *Room Acoustics*. London, U.K.: Elsevier, 1991, ch. 3, p. 56.
[30] D. D. Vries, E. M. Hulsebos, and J. Baan, "Spatial fluctuations in measures for spaciousness," *J. Acoust. Soc. Amer.*, vol. 110, pp. 947–954, Aug. 2001.
[31] X. Pelorson, J. P. Vian, and J. D. Polack, "On the variability of room acoustical parameters: reproducibility and statistical validity," *Appl. Acoust.*, vol. 37, pp. 175–198, 1992.
[32] G. Xuan, W. Zhang, and P. Chai, "EM algorithms of Gaussian mixture model and hidden Markov model," *Proc. IEEE Conf. Image Processing*, vol. 1, pp. 145–148, Oct. 2001.
[33] P. Ramírez, P. Javier, P. Segura, C. José, C. Benítez, C. Carmen, C. de la Torre, and C. Ángel, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, pp. 271–287, Apr. 2004.
[34] I. Potamitis, "Estimation of speech presence probability in the field of microphone array," *IEEE Signal Process. Lett.*, vol. 11, no. 12, pp. 956–959, Dec. 2004.
[35] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. New York: Springer-Verlag, 2001, ch. 2, p. 26.
[36] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.
[37] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*. Berkeley, CA, 1967, pp. 281–297.
[38] C. Elkan, "Using the Triangle Inequality to Accelerate k-Means," in *Proc. 20th Int. Conf. Machine Learning, ICML*, 2003.
[39] Website.. [Online] Available: http://www.30888.com.tw/cars/02rv/savrin/main.htm
[40] T. Pham and B. M. Sadler, "Adaptive wideband aeroacoustic array processing," in *Proc. IEEE Conf. Statistical Signal and Array Processing*, Jun. 1996, pp. 295–298.
[41] M. Friedman and A. Kandel, *Introduction to Pattern Recognition: Statistical, Structural, Neural and Fuzzy Logic Approaches*. Singapore: World Scientific, 1999.

Jwu-Sheng Hu (M'62) was born in Taipei, Taiwan, R.O.C., in 1962. He received the B.S. degree from the Department of Mechanical Engineering, National Taiwan University, Taipei, in 1984, and the M.S. and Ph.D. degrees from the Department of Mechanical Engineering, University of California at Berkeley, in 1988 and 1990, respectively.

He is currently a Professor in the Department of Electrical and Control Engineering, National Chiao-Tung University, Hsinchu. His research interests include microphone array signal processing, active noise control, embedded system design, and robotics.



Chieh-Cheng Cheng was born in 1978. He received the B.S. degree in electrical and control engineering in 2000 from National Chiao Tung University, Taiwan, R.O.C., where he is currently pursuing the Ph.D. degree in the Department of Electrical and Control Engineering.

His research interests include sound source localization, microphone array signal processing, adaptive signal processing, pattern recognition, speech signal processing, and echo and noise cancellation.

Mr. Cheng was the Champion of TI DSP Solutions Design Challenge in 2000 and of the national competition held by Ministry of Education Advisor Office of Taiwan in 2001.

Wei-Han Liu was born in Kaohsiung, Taiwan, R.O.C., in 1977. He received the B.S. and M.S. degrees in electrical and control engineering in 2000 and 2002, respectively, from National Chiao Tung University, Taiwan, where he is currently pursuing the Ph.D. degree in the Department of Electrical and Control Engineering.

His research interests include sound source localization, microphone array signal processing, adaptive signal processing, and speech signal processing, and robot localization.

Mr. Liu was the Champion of the TI DSP Solutions Design Challenge in 2000 and of the national competition held by Ministry of Education Advisor Office of Taiwan in 2001.