PAPER

Single-channel talker localization based on separation of the acoustic transfer function using hidden Markov model and its classification

Ryoichi Takashima[*], Tetsuya Takiguchi[†] and Yasuo Ariki[‡]

Graduate School of System Informatics, Kobe University,
1–1 Rokkodai, Nada-ku, Kobe, 657–8501 Japan

(*Received 1 February 2012, Accepted for publication 1 October 2012*)

Abstract: This paper presents a talker localization method using only a single microphone, where phoneme hidden Markov models (HMMs) of clean speech are introduced to estimate the acoustic transfer function from the user's position. In our previous work, we proposed a Gaussian mixture model (GMM) separation for estimation of the user's position, where the observed speech is separated into the acoustic transfer function and the clean speech GMM. In this paper, we propose an improved method using phoneme HMMs for separation of the acoustic transfer function. This method expresses the speech signal as a network of phoneme HMMs, while our previous method expresses it as a GMM without considering the temporal phonetic changes of the speech signal. The support vector machine (SVM) for classifying the user's position is trained using the separated frame sequences of the acoustic transfer function. Then, for each test data set, the acoustic transfer function is separated, and the position is estimated by discriminating the acoustic transfer function. The effectiveness of this method has been confirmed by talker localization experiments performed in a room environment.

Keywords: Single channel, Talker localization, Acoustic transfer function, Maximum likelihood

PACS number: 43.60.Jn [doi:10.1250/ast.34.176]

1. INTRODUCTION

Many systems using microphone arrays have been developed in an attempt to localize sound sources. Conventional techniques, such as multiple signal classification (MUSIC), cross-power spectrum phase (CSP), and so on (e.g., [1–4]), use simultaneous phase information from microphone arrays to estimate the direction of the arriving signal. There have also been studies on binaural source localization based on interaural differences, such as the interaural level difference and interaural time difference (e.g., [5,6]). However, microphone-array-based systems may not be suitable in some cases because of their size. Therefore, single-channel techniques are of interest, especially in small-device-based scenarios.

The problem of single-microphone source separation is one of the most challenging scenarios in the field of signal processing, and some techniques have been described (e.g., [7–10]). Studies focusing on the techniques for monaural sound source localization are also being carried out [11,12]. In these studies, the information obtained from the external ear, such as head-related transfer functions (HRTFs), is used to localize the sound source.

In our previous work [13], we discussed a sound source localization method using only a single microphone based on discrimination of the acoustic transfer function. In that report, the acoustic transfer function was estimated from observed (reverberant) speech using a clean speech model (speaker-dependent model), where a Gaussian mixture model (GMM) was used to model the features of the clean speech. Using GMM separation, it is possible to estimate the acoustic transfer function using some adaptation data (only several words) uttered from a given position. Because the characteristics of the acoustic transfer function depend on each position, the talker's position, which is trained using some training utterances uttered from the position in advance, can be estimated by discriminating the acoustic transfer function without the external ear if the room environment is the same as that of the training.

If single-channel sound source localization becomes available, it may be possible to apply this technique to devices, such as wearable computers, that are even smaller than existing small devices such as smart phones and some mobile computers, and the performance of some micro-

[*] e-mail: takashima@me.cs.scitec.kobe-u.ac.jp
[†] e-mail: takigu@kobe-u.ac.jp
[‡] e-mail: ariki@kobe-u.ac.jp

phone-array-based systems may also be improved upon combining them with the single-channel technique. In the future, very small sensors might be able to search for people buried under rubble caused by an earthquake by following their voice, for instance.

Our currently proposed method is based on the training of the acoustic transfer function, which is separated from the training speech uttered by the same people from the same position in the same room environment as those for testing. Therefore, there are a number of problems to solve including the changes in the talker, position and room environment. The accurate estimation of the acoustic transfer function from the observed speech is also important for this method. In this paper, we focus on accurate estimation of the acoustic transfer function, and we propose an improved method using phoneme hidden Markov models (HMMs) for separating the acoustic transfer function.

Takiguchi *et al.* [14] proposed an acoustic model adaptation method for distant-talking speech recognition using HMM separation, where the adaptation data were separated into a clean speech model, reverberation model and additive noise model. Using the separated reverberation and noise models, the clean speech HMM was adapted to the test environment and used for the recognition of each test utterance.

In our previous work, the clean speech data was expressed only as a static pattern model (GMM), and this GMM could not express the temporal phonetic changes of clean speech. In addition, as all phonemes are organized into one GMM, the estimated acoustic transfer function was smoothed by the characteristics of all phonemes as a result. In the HMM separation method, on the other hand, as the HMM can deal with a sequential pattern, which has multiple states for each phoneme, phoneme HMMs are able to express more detailed clean speech information including the temporal phonetic changes, and clean speech models of different phonemes do not affect the estimation of the acoustic transfer function. For these reasons, clean speech HMMs can estimate the acoustic transfer function more accurately than a clean speech GMM, and that may enable us to estimate the location of the sound source more accurately.

Unlike the GMM separation in our previous work, however, HMM separation requires texts of the user's utterances in order to estimate the acoustic transfer function. In [14], the utterance texts of adaptation data were given, and the case for unsupervised on-line adaptation was not discussed. In the case of unsupervised adaptation for speech recognition, it is possible for the word recognition results to be used as the utterance text for model adaptation, but our talker localization method requires the utterance text of test data without any dictionaries or knowledge of the language. Therefore, in our method, the observed (reverberant) signal is recognized first using a phoneme recognition system, and the recognition result is used as text information to estimate the acoustic transfer function. This estimation is performed in the cepstral domain employing an approach based on maximum likelihood (ML). This is possible because the cepstral parameters are an effective representation for retaining useful clean speech information. The talker localization experiments discussed in this paper show the effectiveness of the HMM separation of the acoustic transfer function for our single-channel talker localization method.

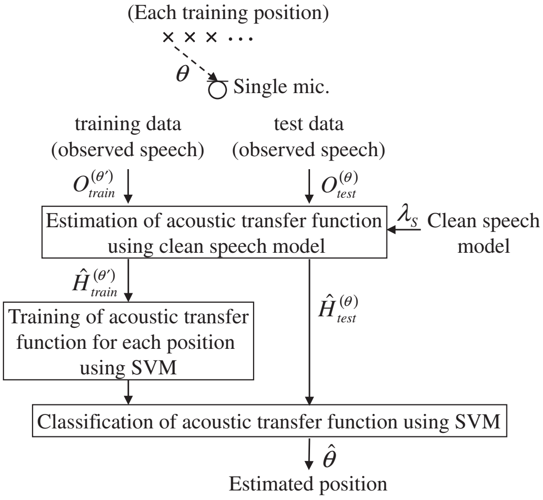
(Each training position)
θ
Single mic.
training data
(observed speech)
test data
(observed speech)
$O_{train}^{(\theta')}$
$O_{test}^{(\theta)}$
Estimation of acoustic transfer function using clean speech model
λ_S
Clean speech model
$\hat{H}_{train}^{(\theta')}$
$\hat{H}_{test}^{(\theta)}$
Training of acoustic transfer function for each position using SVM
Classification of acoustic transfer function using SVM
$\hat{\theta}$
Estimated position

Fig. 1 System overview.

2. ESTIMATION OF THE ACOUSTIC TRANSFER FUNCTION

2.1. System Overview

Figure 1 shows the system overview. First, we record the reverberant speech data $O_{\text{train}}^{(\theta)}$ uttered from each position θ in order to train the acoustic transfer function for θ. Next, the frame sequence of the acoustic transfer function $\hat{H}_{\text{train}}^{(\theta)}$ is estimated from the reverberant speech $O_{\text{train}}^{(\theta)}$ using a clean speech model. This clean speech model is trained using a clean speech database in advance. Then, the support vector machine (SVM) for classifying each user's position θ is trained using the frame sequence of the estimated acoustic transfer function $\hat{H}_{\text{train}}^{(\theta)}$. For test data $O_{\text{test}}^{(\theta)}$ (any utterance), the acoustic transfer function $\hat{H}_{\text{test}}^{(\theta)}$ is estimated in the same way as the training data. The talker's position $\hat{\theta}$ is estimated by discriminating the acoustic transfer function based on the SVM.

Figure 2 shows the estimation process of the acoustic transfer function using the clean speech GMM in our previous method. The acoustic transfer function is estimated by maximizing the likelihood of reverberant speech

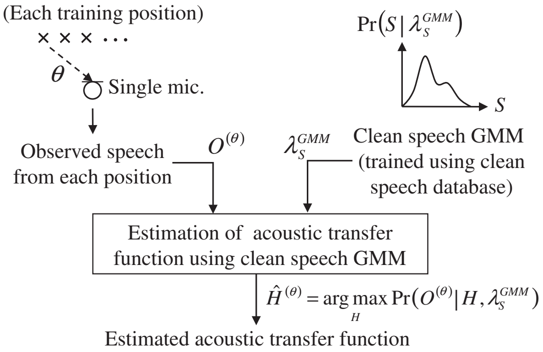
(Each training position)
θ
Single mic.
Observed speech from each position
$O^{(\theta)}$
λ_S^{GMM}
$\Pr(S|\lambda_S^{GMM})$
S
Clean speech GMM (trained using clean speech database)
Estimation of acoustic transfer function using clean speech GMM
$\hat{H}^{(\theta)} = \arg\max_H \Pr(O^{(\theta)}|H,\lambda_S^{GMM})$
Estimated acoustic transfer function

Fig. 2 Estimation of the acoustic transfer function using a GMM of clean speech.

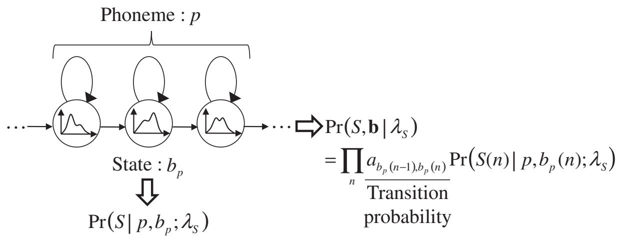
Phoneme : p
State : b_p
$\Pr(S|p,b_p;\lambda_S)$
$\Pr(S,\mathbf{b}|\lambda_S)$
$=\prod_n a_{b_p(n-1),b_p(n)} \Pr(S(n)|p,b_p(n);\lambda_S)$
Transition probability

Fig. 3 Clean speech model expressed by an HMM.

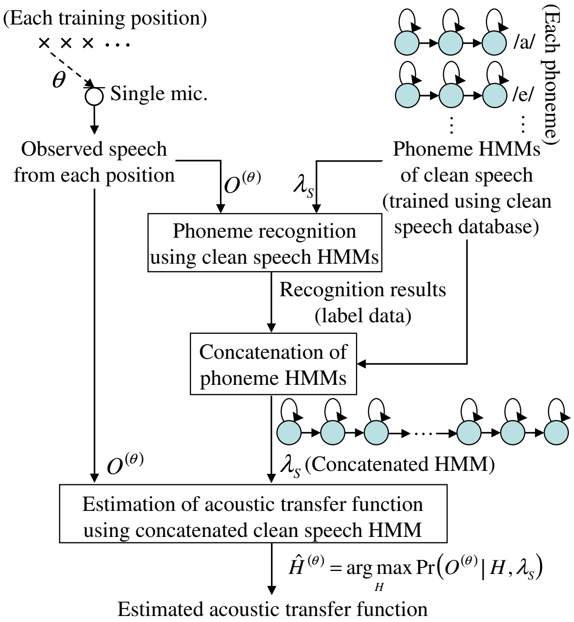
(Each training position)
θ
Single mic.
Observed speech from each position
$O^{(\theta)}$
λ_S
/a/
/e/
(Each phoneme)
Phoneme HMMs of clean speech (trained using clean speech database)
Phoneme recognition using clean speech HMMs
Recognition results (label data)
Concatenation of phoneme HMMs
$O^{(\theta)}$
λ_S (Concatenated HMM)
Estimation of acoustic transfer function using concatenated clean speech HMM
$\hat{H}^{(\theta)} = \arg\max_H \Pr(O^{(\theta)}|H,\lambda_S)$
Estimated acoustic transfer function

Fig. 4 Estimation of the acoustic transfer function using phoneme HMMs of clean speech.

using a clean speech GMM. In this method, the clean speech data is expressed as only a static pattern model, and this GMM cannot express the temporal phonetic changes of clean speech. In addition, all phonemes are organized into one GMM. This causes the smoothing of the estimated acoustic transfer function by the characteristics of all phonemes.

In our proposed method, on the other hand, the frame sequence of the acoustic transfer function $\hat{H}^{(\theta)}$ is estimated using clean speech HMMs. Figure 3 shows an example of a clean speech HMM. An HMM is a state transition model, and each state has a GMM as the posterior probability. The likelihood is calculated as the product of the transition probability $a_{b_p(n-1),b_p(n)}$ and the posterior probability for each state of the frame n. This method expresses the clean speech as a network of phoneme HMMs, where the phoneme HMMs express more detailed clean speech information including the temporal phonetic changes, and clean speech models of different phonemes do not affect the estimation of the acoustic transfer function.

Figure 4 shows the estimation process of the acoustic transfer function using the clean speech HMMs in our proposed method. As the clean speech HMMs are trained for each phoneme, our proposed method requires texts of the user's utterances in order to construct the network of phoneme HMMs. For this purpose, the phoneme sequence of the reverberant speech data is first recognized by using each phoneme HMM of clean speech data. Using the recognition results (1-best hypothesis), the phoneme HMMs are concatenated, and the frame sequence of the acoustic transfer function $\hat{H}^{(\theta)}$ is estimated from the reverberant speech $O^{(\theta)}$ based on an ML estimation approach using the concatenated HMM.

2.2. Cepstrum Representation of Reverberant Speech

The reverberant speech signal $o(t)$ in a room environment is generally considered to be the convolution of clean speech and the acoustic transfer function $o(t) = \sum_{l=0}^{L-1} s(t-l)h(l)$, where $s(t)$, $h(l)$ and L are a clean speech signal, an acoustic transfer function (room impulse response) from the sound source to the microphone and the length of the acoustic transfer function, respectively.

In recent studies on robust speech recognition and speech dereverberation, the reverberant speech in the short-term Fourier transform (STFT) domain is often modeled so that each frequency bin of the reverberant speech is represented by the convolution of the frame sequences of clean speech and the acoustic transfer function as follows [15,16].

$$O_{\mathrm{spc}}(\omega;n) = \sum_{l'=0}^{L'-1} S_{\mathrm{spc}}(\omega;n-l') \cdot H_{\mathrm{spc}}(\omega;l') \qquad (1)$$

Here, $O_{\mathrm{spc}}(\omega;n)$, $S_{\mathrm{spc}}(\omega;n)$ and $H_{\mathrm{spc}}(\omega;n)$ are the ωth frequency bins of short-term linear spectra of frame n. L' is the length of the acoustic transfer function in the STFT domain. However, such modeling is complex for estimating the frame sequence of the acoustic transfer function, and it is difficult to deal with the estimated components of the acoustic transfer function for this talker localization task. Therefore, in this paper, we employ a simpler model of the reverberant speech, which is approximately represented as the product of clean speech and the acoustic transfer function.

$$O_{\mathrm{spc}}(\omega; n) \approx S_{\mathrm{spc}}(\omega; n) \cdot H_{\mathrm{spc}}(\omega; n) \tag{2}$$

Cepstral parameters are an effective representation to retain useful speech information in speech recognition. Therefore, we use the cepstrum for acoustic modeling that is necessary to estimate the acoustic transfer function. The cepstrum of the reverberant speech is given by the inverse Fourier transform of the log spectrum.

$$O_{\mathrm{cep}}(d; n) \approx S_{\mathrm{cep}}(d; n) + H_{\mathrm{cep}}(d; n), \tag{3}$$

where O_{cep}, S_{cep} and H_{cep} are cepstra for the reverberant speech signal, clean speech signal and acoustic transfer function, respectively. d is the dimension of the cepstrum. As shown in Eq. (3), if O and S are observed, H can be obtained by

$$H_{\mathrm{cep}}(d; n) \approx O_{\mathrm{cep}}(d; n) - S_{\mathrm{cep}}(d; n). \tag{4}$$

However, S_{cep} cannot actually be observed. Therefore, H_{cep} is estimated by maximizing the likelihood of reverberant speech using a clean speech HMM. From the following section, the cepstral variables O_{cep}, S_{cep} and H_{cep} are written as O, S and H for simplicity, respectively.

2.3. Maximum-Likelihood-Based Parameter Estimation

This section presents a new method for estimating the acoustic transfer function. The estimation is implemented by maximizing the likelihood of the training data from the user's position. In [17], an ML estimation method for decreasing the acoustic mismatch for a telephone channel was described, and in [18] channel distortion and noise were simultaneously estimated using an expectation maximization (EM) method.

The frame sequence of the acoustic transfer function in Eq. (4) is estimated in an ML manner by using the EM algorithm, which maximizes the likelihood of the observed speech:

$$\hat{H} = \operatorname*{argmax}_{H} \Pr(O|H, \lambda_S). \tag{5}$$

Here, λ_S denotes the set of concatenated clean speech HMM parameters, while the suffix S represents the clean speech in the cepstral domain. The EM algorithm is a two-step iterative procedure. In the first step, called the expectation step, the following auxiliary function is computed.

$$\begin{aligned} Q(\hat{H}|H) &= E[\log \Pr(O, p, b_p, c_p|\hat{H}, \lambda_S)|H, \lambda_S] \\ &= \sum_p \sum_{b_p} \sum_{c_p} \frac{\Pr(O, p, b_p, c_p|H, \lambda_S)}{\Pr(O|H, \lambda_S)} \\ &\quad \cdot \log \Pr(O, p, b_p, c_p|\hat{H}, \lambda_S) \end{aligned} \tag{6}$$

Here, b_p and c_p represent the unobserved state sequence and the unobserved mixture component labels corresponding to the phoneme p in the observation sequence O, respectively.

The joint probability of observing sequences O, b and c can be calculated as

$$\begin{aligned} &\Pr(O, p, b_p, c_p|\hat{H}, \lambda_S) \\ &\quad = \prod_n a_{b_p(n-1), b_p(n)} w_{b_p(n), c_p(n)} \\ &\quad\quad \cdot \Pr(O(n)|p, b_p(n), c_p(n); \hat{H}, \lambda_S), \end{aligned} \tag{7}$$

where n, a and w represent the frame, the transition probability and the mixture weight, respectively. Since we consider the acoustic transfer function as additive noise in the cepstral domain, the mean of the mixture k of state j in the model λ_O is derived by adding the acoustic transfer function. Therefore, Eq. (7) can be written as

$$\begin{aligned} &\Pr(O, p, b_p, c_p|\hat{H}, \lambda_S) \\ &\quad = \prod_n a_{b_p(n-1), b_p(n)} w_{b_p(n), c_p(n)} \\ &\quad\quad \cdot N(O(n); \mu^{(S)}_{p,j,k} + \hat{H}(n), \Sigma^{(S)}_{p,j,k}), \end{aligned} \tag{8}$$

where $N(O; \mu, \Sigma)$ denotes the multivariate Gaussian distribution. The following derivation is straightforward [19]:

$$\begin{aligned} &Q(\hat{H}|H) \\ &= \sum_p \sum_i \sum_j \sum_n \\ &\quad \Pr(O(n), p, b_p(n) = j, b_p(n-1) = i|H, \lambda_S) \log a_{p,i,j} \\ &+ \sum_p \sum_j \sum_k \sum_n \\ &\quad \Pr(O(n), p, b_p(n) = j, c_p(n) = k|H, \lambda_S) \log w_{p,j,k} \\ &+ \sum_p \sum_j \sum_k \sum_n \\ &\quad \Pr(O(n), p, b_p(n) = j, c_p(n) = k|H, \lambda_S) \\ &\quad \cdot \log N(O(n); \mu^{(S)}_{p,j,k} + \hat{H}(n), \Sigma^{(S)}_{p,j,k}). \end{aligned} \tag{9}$$

Here, $\mu^{(S)}_{p,j,k}$ and $\Sigma^{(S)}_{p,j,k}$ are the mean vector and the (diagonal) covariance matrix in the concatenated clean speech HMM, respectively. i is the state of the previous frame. It is possible to train these parameters by using a clean speech database.

Next, we focus only on the term involving H.

$$\begin{aligned} &Q(\hat{H}|H) \\ &\quad = -\sum_p \sum_j \sum_k \sum_n \gamma_{p,j,k}(n) \end{aligned}$$

$$- \sum_{d=1}^{D} \left\{ \frac{1}{2} \log(2\pi)^D \sigma_{p,j,k,d}^{(S)^2} + \frac{(O(d;n) - \mu_{p,j,k,d}^{(S)} - \hat{H}(d;n))^2}{2\sigma_{p,j,k,d}^{(S)^2}} \right\} \quad (10)$$

$$\gamma_{p,j,k}(n) = \Pr(O(n), p, j, k | H, \lambda_S) \quad (11)$$

Here, D is the dimension of the observation vector $O(n)$, and $\mu_{p,j,k,d}^{(S)}$ and $\sigma_{p,j,k,d}^{(S)^2}$ are the dth dimension of the mean value and that of the diagonal variance value, respectively.

The maximization step in the EM algorithm becomes max $Q(\hat{H}|H)$. The re-estimation formula can therefore be derived, knowing that $\partial Q(\hat{H}|H)/\partial \hat{H} = 0$ as

$$\hat{H}(d;n) = \frac{\displaystyle\sum_p \sum_j \sum_k \gamma_{p,j,k}(n) \frac{O(d;n) - \mu_{p,j,k,d}^{(S)}}{\sigma_{p,j,k,d}^{(S)^2}}}{\displaystyle\sum_p \sum_j \sum_k \frac{\gamma_{p,j,k}(n)}{\sigma_{p,j,k,d}^{(S)^2}}}. \quad (12)$$

After calculating the acoustic transfer function for all the training data, the SVM for classifying each user's position is trained using the estimated acoustic transfer function. The SVM is a classifier that has high generalization capability and robustness to outliers, and it tends to show relatively high performance even if the amount of training data is small. This is why the SVM may be suitable for our proposed method, for which a small number of user's training utterances for each position is preferred and the estimated acoustic transfer function (which may include some misestimation) is used. For test data, the acoustic transfer function is estimated in the same way as the training data using a label sequence obtained from the phoneme recognition system. The talker's position is estimated by discriminating the acoustic transfer function based on the SVM.

3. EXPERIMENTS

3.1. Simulation Experimental Conditions

The new talker localization method was evaluated in both a simulated reverberant environment and a real environment. In the simulated environment, the reverberant speech was simulated using a linear convolution of clean speech and the impulse response. The impulse response was taken from the RWCP database in real acoustical environments [20]. Figure 5 shows the experimental room environment. The size of the recording room was about 6.7 m × 4.2 m (width × depth), and the reverberation time was 300 ms. A loudspeaker was located on a circular arc with a radius of 2,020 mm. The microphone was located 420 mm from the center of the circle. Therefore, the distances from the microphone to every loudspeaker position were from 1,600 to 1,900 mm. The height of the loudspeaker at every position was 1,720 mm, and that of the microphone was 1,620 mm. For each position, the loudspeaker faced the microphone. Figure 6 shows the impulse response (90 degrees) [13].

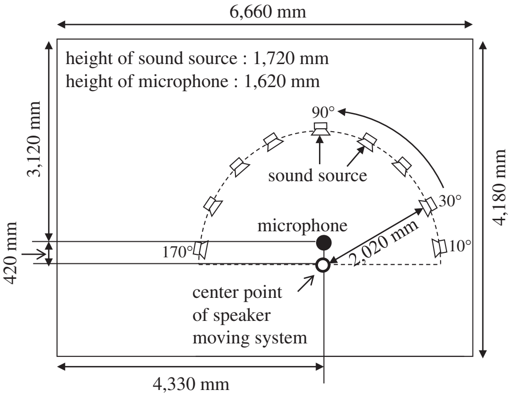

Fig. 5 Experimental room environment for simulation.

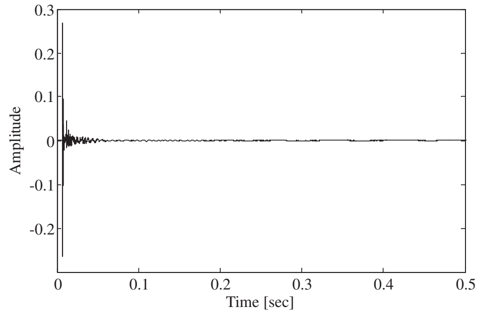

Fig. 6 Impulse response (90 degrees, reverberation time: 300 ms).

The speech signal was sampled at 12 kHz and windowed with a 32 ms Hamming window every 8 ms. The experiment utilized the speech data of five males in the ATR Japanese speech database. The clean speech HMM (speaker-dependent model) was trained using 2,620 words. Each phoneme HMM was a simple left-right model having three states with self-transitions, and each state has 32 Gaussian mixture components. The total number of phonemes is 54. The test data for one location consisted of 1,000 words, and 16-order mel-frequency cepstral coefficients (MFCCs) were used as feature vectors. The total number of test data for one location was 1,000 (words) × 5 (males). The number of data used to train the acoustic transfer function for one location was 10, 20, 30, 40 or 50 words. The speech data for training the clean speech model, training the acoustic transfer function and testing were spoken by the same person but had different text utterances. The speaker positions for training and testing consisted of three positions (30, 90 and 130 degrees), five positions (10, 50, 90, 130 and 170 degrees), seven positions (30, 50, 70, ..., 130 and 150 degrees) and nine positions (10, 30, 50, 70, ..., 150 and 170 degrees).

For each test data (word), the position is classified using the SVM. We used SVM^{light} [21] as an SVM with an RBF (Gaussian) kernel. Then, the SVM was extended using the one-vs-rest method in order to carry out multiclass classification. These experiments were carried out for each speaker, and the localization accuracy was averaged for the five speakers.

3.2. Experimental Results in a Simulated Reverberant Environment

We compared the following methods for the discrimination of each position.

HMM (1-best hypothesis) The proposed method using the acoustic transfer function estimated using clean speech HMMs. The phoneme HMMs are concatenated using the phoneme recognition results (1-best hypothesis).

HMM (correct transcription) The proposed method using the acoustic transfer function estimated using clean speech HMMs. The phoneme HMMs are concatenated using the correct transcription.

GMM Our previous method using the acoustic transfer function estimated using a clean speech GMM. The clean speech GMM was trained using the same 2,620 words as those for the HMMs and has 64 Gaussian mixture components.

Observed speech While the above three methods localize the sound source by discriminating the acoustic transfer function, this method localizes the sound source by directly discriminating the observed speech without separating the acoustic transfer function. The observed speech includes not only the acoustic transfer function but also clean speech, which is meaningless information for sound source localization.

For these methods, the talker's positions are discriminated using the SVM. The experimental conditions, such as feature vectors and the number of training data, are the same in each method.

Table 1 shows the mean square error (MSE) of the acoustic transfer function separated using a clean speech GMM, clean speech HMMs with the 1-best hypothesis, and HMMs with the correct transcription. The MSE is the mean value of the square error for each frame calculated using the following equation:

$$\mathrm{MSE} = \frac{1}{N}\sum_{n}\sum_{d}(H_{\mathrm{true}}(d;n) - \hat{H}(d;n))^2, \qquad (13)$$

where the acoustic transfer function calculated by Eq. (4) using the true clean speech data is used as the ground truth $H_{\mathrm{true}}(d;n)$. Figure 7 shows the variance of each cepstral order of H_{true}. The MSE of the observed speech in Table 1 shows the mean Euclidean distance between the true acoustic transfer function and the observed speech, which was calculated by substituting $O(d;n)$ for $\hat{H}(d;n)$ in Eq. (13). As shown in this table, the MSE of the acoustic transfer function that was estimated using clean speech HMMs was smaller than that estimated using a clean speech GMM. This means that the proposed method can estimate the acoustic transfer function more accurately than our previous method. In addition, when the correct transcription was used instead of the 1-best hypothesis for concatenating the phoneme HMMs, the MSE decreased even more.

Table 1 Mean square error of the acoustic transfer function separated using a clean speech GMM, clean speech HMMs with the 1-best hypothesis and HMMs with the correct transcription. The MSE of the observed speech was calculated by substituting $O(d;n)$ for $\hat{H}(d;n)$ in Eq. (13).

	Observed speech	GMM	HMM (1-best hypothesis)	HMM (correct transcription)
MSE	9485.97	2264.33	2096.14	1968.36

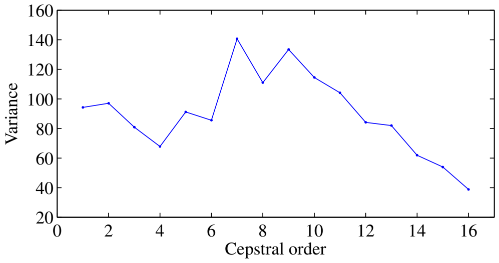

Fig. 7 Variance of each cepstral order of H_{true}.

Figure 8 shows the mel spectra of the ground truth of the acoustic transfer function, estimated acoustic transfer functions, and the observed speech of a sample frame. The 32-order mel spectrum was obtained by computing the inverse cosine transform of the 32-order MFCCs, where the estimated 16-order MFCCs were extended to 32-order MFCCs using zero padding. Then, the mel spectra had normalized energies because the 0th dimension of the MFCCs (which is equivalent to the energy of the mel spectrum and is not discussed in this study) was also padded with zeros. This figure also shows that the clean speech HMMs suppressed the influence of clean speech more effectively and estimated the acoustic transfer function more correctly than the clean speech GMM.

Table 2 shows a comparison of the four methods for each number of training data, where the number of positions was three. Table 3 shows a comparison for each number of positions, where the number of training data was 50 words. The proposed method and our previous method showed higher accuracies than the use of the observed

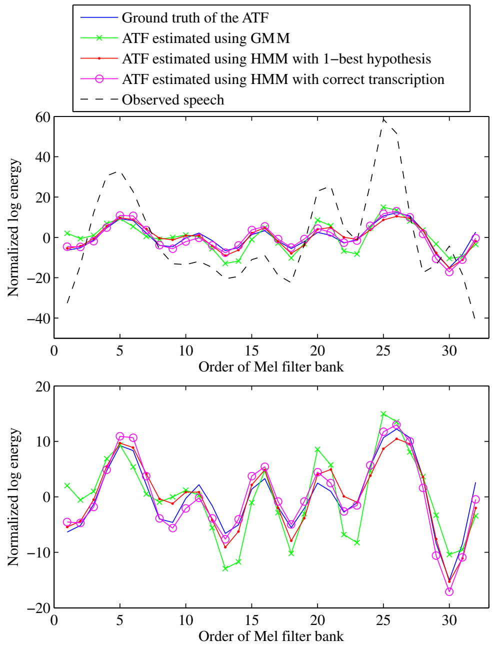

Fig. 8 Mel spectra of the ground truth of the acoustic transfer function (ATF), estimated acoustic transfer functions and observed speech of a sample frame (top figure). The bottom figure is a close-up of the estimated acoustic transfer functions.

Table 2 Localization accuracies [%] of compared methods for each number of training data (words), where the number of positions was three. The method of **Observed speech** localizes the sound source by discriminating the observed speech directly, while the other three methods discriminate the separated acoustic transfer function. The parenthetic numbers show the differences from the accuracy of the HMMs (1-best hypothesis) [%].

Number of training data (words)	50	40	30	20	10
HMM (1-best hypothesis)	82.9	82.4	82.3	80.6	79.9
HMM (correct transcription)	83.9 (0.9)	83.3 (0.9)	83.0 (0.8)	81.0 (0.4)	79.5 (−0.4)
GMM	80.5 (−2.4)	80.2 (−2.2)	79.2 (−3.1)	76.9 (−3.6)	75.1 (−4.8)
Observed speech	53.2 (−29.7)	50.2 (−32.2)	46.6 (−35.7)	40.7 (−39.9)	35.7 (−44.2)

Table 3 Localization accuracies [%] of compared methods for each number of positions, where the number of training data was 50 words. The parenthetic numbers show the differences from the accuracy of the HMMs (1-best hypothesis) [%].

Number of positions	3	5	7	9
HMM (1-best hypothesis)	82.9	61.4	56.0	46.6
HMM (correct transcription)	83.9 (0.9)	62.0 (0.7)	58.0 (1.9)	48.0 (1.5)
GMM	80.5 (−2.4)	57.0 (−4.4)	53.6 (−2.5)	44.0 (−2.5)
Observed speech	53.2 (−29.7)	27.0 (−34.4)	30.3 (−25.7)	22.5 (−24.1)

speech. Our methods separate the acoustic transfer function from the observed speech signal; thus, the use of the estimated acoustic transfer function will not be affected greatly by the characteristics of the clean speech (phonemes). In the comparison between the proposed method (HMMs with 1-best hypothesis) and our previous method (GMM), the newly proposed method showed higher performances than our previous method by an accuracy of at least 2.2% for every set of conditions, and these differences in the accuracy were significant ($P < 0.01$, chi-square test).

As shown in Table 2, the differences between the accuracy of the proposed method and our previous method or the method using the observed speech became larger as the number of training data decreased. In the results of the use of the observed speech, as the number of training data decreased, the classification boundaries were biased by the utterance contents of the training data (i.e., overtraining). These biases also arose in the results of the proposed method and our previous method, because the clean speech components are not removed from the observed signal completely, and remained in the estimated acoustic transfer function. However, the clean speech HMMs could suppress the influence of the difference in the utterance texts of training and testing more effectively than the clean speech GMM; thus, the degree of bias was reduced in the proposed method. This is why the differences in accuracy became larger when a small number of training data was used.

In the comparison between the use of the 1-best hypothesis and the correct transcription, there were no significant differences in the accuracy ($P \geq 0.03$) for each number of training data, when the number of positions was three (shown in Table 2), although the use of the correct transcription could estimate the acoustic transfer function more accurately than the use of the 1-best hypothesis. However, as shown in Table 3, there were significant differences in the accuracy ($P < 0.01$) when the number of positions was seven or nine. Figures 9 and 10 show the 7th and 10th orders of the mel-cepstral coefficients, which had the highest ratios of the within-class variances to the between-class variances (Fisher's ratios), of mean acoustic transfer function values for each word in the case of three and seven positions, respectively. The acoustic transfer functions are calculated by Eq. (4).

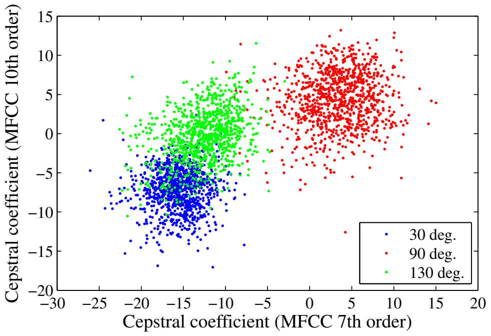

Fig. 9 Mean acoustic transfer function values for three positions.

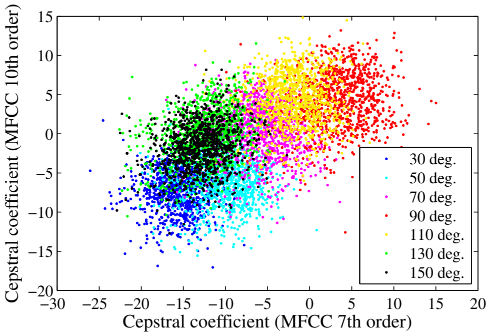

Fig. 10 Mean acoustic transfer function values for seven positions.

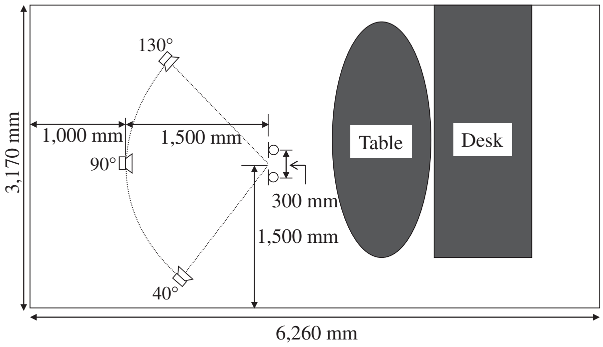

Fig. 11 Experimental room environment and the loudspeaker position.

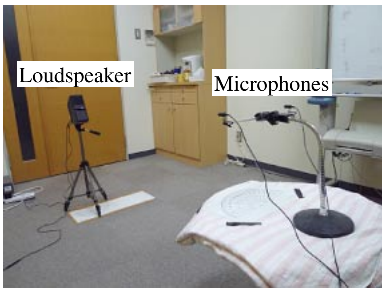

Fig. 12 Photograph of the recording environment.

As shown in these figures, when the number of positions is three, the distribution of the acoustic transfer function for each position can be discriminated relatively easily. However, when the number of positions is seven, it is difficult to discriminate the distribution for each position. Therefore, the accurate estimation of the acoustic transfer function may contribute more to the discrimination of each position as the number of positions increases. This is why the differences in accuracy between the use of the 1-best hypothesis and that of the correct transcription were not significant when the number of positions was small, but the differences were significant when the number of positions was large.

3.3. Experimental Results in a Real Environment

The proposed method was also evaluated in a real environment. Figure 11 shows the experimental room environment and the position of the loudspeaker. Figure 12 depicts the recording environment. The size of the recording room was about 6.3 m × 3.2 m × 2.8 m (width × depth × height). The reverberation time was about 350 ms, and the SNR was about 41.49 [dB]. The distance from each position to the microphone was about 1.5 m. The speech signal was recorded using two microphones in order to provide a comparison with conventional CSP analysis, but the signal recorded by only one of the microphones was used for the proposed method. The microphone was a directional type (Sony ECM-66B). There were three loudspeaker positions (40, 90 and 130 degrees) for training and testing, and one loudspeaker (BOSE Mediamate II) was used for each position.

The experiment utilized the speech data uttered by a male in the ATR Japanese speech database. The clean speech HMM was trained using 2,620 words. We recorded 216 words for each location. Then, 50 of these words were used to train the acoustic transfer function for one location, and the other 166 words were used for the test data for the location. The estimation accuracy was calculated by 4-fold cross-validation. The total number of test data was 648 words (216 words × 3 positions). The speech data for training the clean speech model, training the acoustic transfer function and testing were spoken by the same person but had different text utterances. The other experimental conditions are the same as those described in preceding sections.

The proposed method was compared with a conventional CSP algorithm [2] based on two microphones. We evaluated the performances of these methods under several testing conditions, where the orientation or location of the loudspeaker changed from that of the loudspeaker for training, or both of them matched those for training. Figure 13 shows the differences in the orientation and position of the loudspeaker. The orientation for testing was changed to 0 (matched training condition), 45 and 90

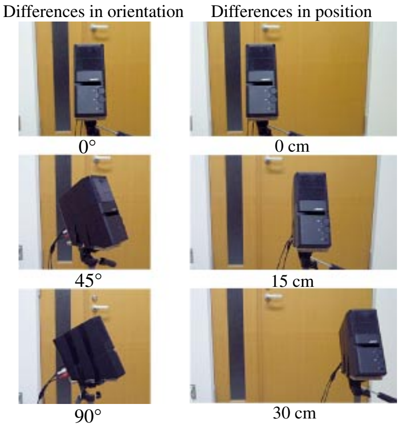

Fig. 13 Differences in orientation and position of the loudspeaker.

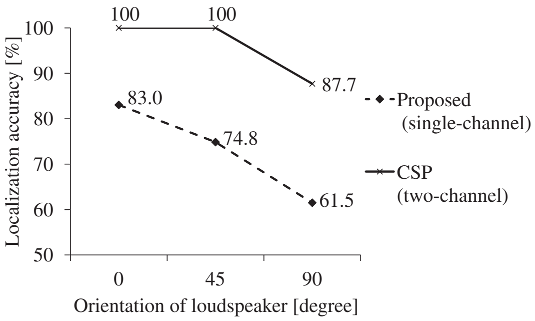

Fig. 14 Comparison of performance for each orientation of the loudspeaker for testing.

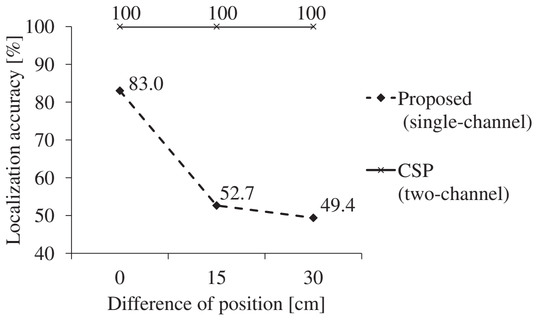

Fig. 15 Comparison of performance for each difference between the positions of the loudspeaker for training and testing.

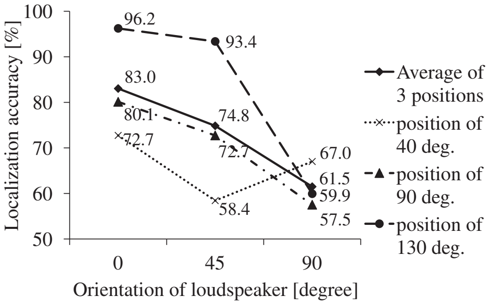

Fig. 16 Localization accuracy of each position and orientation of the loudspeaker for testing.

degrees, and the position was changed to 0 (matched training condition), 15 and 30 cm.

Figures 14 and 15 show comparisons of the performances for each difference in the orientation and position of the loudspeaker between those for training and testing, respectively. As shown in Fig. 14, the CSP algorithm could estimate the location with an accuracy of 100% except for the case where the orientation of the loudspeaker was 90 degrees. The reason why the performance of the CSP algorithm degraded when the orientation was 90 degrees may be that the effect of the reflected waves from the wall (reverberation) became larger.

On the other hand, the localization accuracy of the proposed method degraded as the orientation angle of the loudspeaker changed significantly. This means that the acoustic transfer function depends on not only the position but also the orientation of the speaker, and the characteristics of the acoustic transfer function changed from those for training despite being measured from the same position. However, if the acoustic transfer function of each orientation is trained, the proposed system may be able to estimate not only the position but also the orientation of the talker's head, and conventional microphone array systems may be able to estimate the position more accurately in combination with the proposed system even when the talker is facing away from the microphone.

As shown in Fig. 15, the accuracy degraded drastically at the point where the difference between the positions of the loudspeaker for training and testing was 15 cm, while the CSP algorithm estimated the location with an accuracy of 100% for every condition. This means the characteristics of the acoustic transfer function changed drastically when the position was changed by 15 cm, although the phase difference used in the CSP algorithm changed little.

Figures 16 and 17 show the localization accuracies of each position of the sound source in these experiments. As shown in Fig. 16, the localization accuracy of the 130 degree position was drastically degraded by changing the orientation from 45 to 90 degrees. That was because when the orientation was 90 degrees, the loudspeaker at 130 degrees emitted utterances toward the wall close to the loudspeaker, and the acoustic transfer function might have changed greatly.

As shown in Fig. 17, the degradation in the accuracy at the 40 degree position was the most drastic among the three positions. This might be because the 40 degree training position was the closest to the wall among the three loudspeaker positions, and the estimation of that position was the most sensitive to changes in the position during

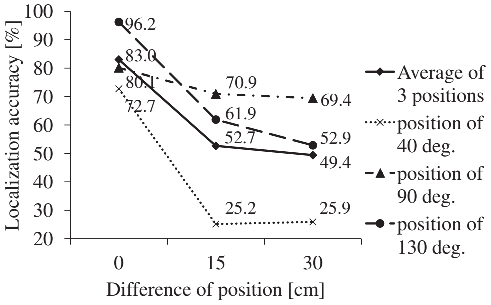

Fig. 17 Localization accuracy of each position as a function of the difference between the positions of the loudspeaker for training and testing.

testing. On the other hand, the degradation in the accuracy at the 90 degree position was relatively small compared with other positions. This might be because the 90 degree position was far from the wall compared with other positions, and the estimation of that position was relatively robust to changes in the position. In order to increase the robustness to changes in the testing environment, feature selection or feature transformation techniques may need to be applied to reduce the sensitivity of the acoustic transfer function.

4. CONCLUSION

This paper has described a sound source (talker) localization method using a single microphone based on the discrimination of the acoustic transfer function. In order to improve the performance of our previous method, phoneme hidden Markov models (HMMs) of clean speech are introduced to estimate the acoustic transfer function from the user's position more accurately. The acoustic transfer function is estimated using phoneme HMMs of clean speech and a label sequence obtained from phoneme recognition. In comparative experiments in a room environment, the proposed method could estimate the acoustic transfer function more accurately than our previous method, and it improved the localization accuracy of our previous method by at least 2.2% for every experimental condition. The localization accuracy decreases as the number of positions increases. In order to localize the talker more accurately under conditions where there is a large number of positions, our method needs to estimate the acoustic transfer function more accurately. To achieve this, estimation of more appropriate test data utterance texts is also needed.

The localization accuracy of the proposed method degraded as the orientation angle of the loudspeaker changed because the acoustic transfer function depends not only on the position but also on the orientation of the speaker. However, if the acoustic transfer function of each orientation is trained, the proposed system may be able to estimate not only the position but the orientation of the talker's head. Information about the talker's head orientation may also be important, especially in multiuser conversation scenarios, such as robot communication scenarios, because it can determine not only who is talking but also who he/she is talking to. In addition, conventional microphone array systems may be able to estimate the position more accurately in combination with the proposed system even when the talker is facing away from the microphone.

However, the proposed approach was too sensitive to changes in the testing environment. As described in Sect. 1, there are a number of problems that need to be solved. Therefore, we will study model adaptation techniques, feature selection and feature transformation techniques in order to increase the method's robustness to changes in the talker and other environmental conditions. Future work will also include estimating the speaker's position which has not been pre-trained by investigating on-line training or adaptation techniques.

ACKNOWLEDGMENT

This work was supported by a Grant-in-Aid for JSPS Fellows (23·2495).

REFERENCES

[1] D. Johnson and D. Dudgeon, *Array Signal Processing* (Prentice Hall, Upper Saddle River, NJ, 1996).

[2] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using CSP analysis," *Proc. ICASSP 96*, Vol. 2, pp. 921–924 (1996).

[3] F. Asano, H. Asoh and T. Matsui, "Sound source localization and separation in near field," *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, **E83-A**, 2286–2294 (2000).

[4] Y. Denda, T. Nishiura and Y. Yamashita, "Robust talker direction estimation based on weighted CSP analysis and maximum likelihood estimation," *IEICE Trans. Inf. Syst.*, **E89-D**, 1050–1057 (2000).

[5] F. Keyrouz, Y. Naous and K. Diepold, "A new method for binaural 3-D localization based on HRTFs," *Proc. ICASSP 06*, Vol. 5, pp. 341–344 (2006).

[6] M. Takimoto, T. Nishino and K. Takeda, "Estimation of a talker and listener's positions in a car using binaural signals," *Proc. 4th Jt. Meet. ASA and ASJ* (*ASA/ASJ 06*), p. 3216 (2006).

[7] T. Kristjansson, H. Attias and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," *Proc. ICASSP 04*, Vol. 2, pp. 817–820 (2004).

[8] B. Raj, M. V. S. Shashanka and P. Smaragdis, "Latent direchlet decomposition for single channel speaker separation," *Proc. ICASSP 06*, Vol. 5, pp. 821–824 (2006).

[9] G.-J. Jang, T.-W. Lee and Y.-H. Oh, "A subspace approach to single channel signal separation using maximum likelihood weighting filters," *Proc. ICASSP 03*, Vol. 5, pp. 45–48 (2003).

[10] T. Nakatani, B.-H. Juang, K. Kinoshita and M. Miyoshi, "Speech dereverberation based on probabilistic models of source and room acoustics," *Proc. ICASSP 06*, Vol. 1, pp. 821–824 (2006).

[11] A. Fuchs, C. Feldbauer and M. Stark, "Monaural sound

localization," *Proc. Interspeech 2011*, pp. 2521–2524 (2011).

[12] R. Kliper, H. Kayser, D. Weinshall, I. Nelken and J. Anemuller, "Monaural azimuth localization using spectral dynamics of speech," *Proc. Interspeech 2011*, pp. 33–36 (2011).

[13] T. Takiguchi, Y. Sumida, R. Takashima and Y. Ariki, "Single-channel talker localization based on discrimination of acoustic transfer functions," *EURASIP J. Adv. Signal Process. 2009*, pp. 9 (2009).

[14] T. Takiguchi, S. Nakamura and K. Shikano, "HMM-separation-based speech recognition for a distant moving speaker," *IEEE Trans. Speech Audio Process.*, **9**, 127–140 (2001).

[15] A. Sehr, R. Maas and W. Kellermann, "Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, **18**, 1676–1691 (2010).

[16] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," *Proc. ICASSP 08*, pp. 85–88 (2008).

[17] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Process.*, **4**, 190–202 (1996).

[18] T. Kristiansson, B. Frey, L. Deng and A. Acero, "Joint estimation of noise and channel distortion in a generalized EM framework," *Proc. IEEE Automatic Speech Recognition and Understanding Workshop* (*ASRU01*), pp. 155–158 (2001).

[19] B.-H. Juang, "Maximum-likelihood estimation of mixture multivariate stochastic observations of Markov chains," *AT&T Tech. J.*, **64**, 1235–1249 (1985).

[20] S. Nakamura, "Acoustic sound database collected for hands-free speech recognition and sound scene understanding," *Proc. Int. Workshop Hands-Free Speech Communication* (*HSC01*), pp. 43–46 (2001).

[21] T. Joachims, *Making Large-scale SVM Learning Practical* (MIT Press, Cambridge, Mass., 1999).

Ryoichi Takashima received his B.E. and M.E. degrees in computer science from Kobe University in 2008 and 2010, respectively. His current research interests include speech signal processing and pattern recognition. He is a member of ASJ.

Tetsuya Takiguchi received the B.S. degree in applied mathematics from Okayama University of Science, Okayama, Japan, in 1994, and the M.E. and Dr. Eng. degrees in information science from Nara Institute of Science and Technology, Nara, Japan, in 1996 and 1999, respectively. From 1999 to 2004, he was a researcher at IBM Research, Tokyo Research Laboratory, Kanagawa, Japan. He is currently an Associate Professor at Kobe University. His research interests include statistic signal processing and pattern recognition. He received the Awaya Award from the Acoustical Society of Japan in 2002. He is a member of the IEEE, the IPSJ, and the ASJ.

Yasuo Ariki received his B.E., M.E. and Ph.D. in information science from Kyoto University in 1974, 1976 and 1979, respectively. He was an assistant professor at Kyoto University from 1980 to 1990, and stayed at Edinburgh University as visiting academic from 1987 to 1990. From 1990 to 1992 he was an associate professor and from 1992 to 2003 a professor at Ryukoku University. Since 2003 he has been a professor at Kobe University. He is mainly engaged in speech and image recognition and interested in information retrieval and database. He is a member of IEEE, IPSJ, JSAI, ITE and IIEEJ.