Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 918404, 9 pages
doi:10.1155/2009/918404

Research Article

Single-Channel Talker Localization Based on Discrimination of Acoustic Transfer Functions

Tetsuya Takiguchi, Yuji Sumida, Ryoichi Takashima, and Yasuo Ariki

Organization of Advanced Science and Technology, Kobe University, Kobe 657-8501, Japan

Correspondence should be addressed to Tetsuya Takiguchi, takigu@kobe-u.ac.jp

Received 5 June 2008; Revised 3 November 2008; Accepted 5 February 2009

Recommended by Aggelos Pikrakis

This paper presents a sound source (talker) localization method using only a single microphone, where a Gaussian Mixture Model (GMM) of clean speech is introduced to estimate the acoustic transfer function from a user's position. The new method is able to carry out this estimation without measuring impulse responses. The frame sequence of the acoustic transfer function is estimated by maximizing the likelihood of training data uttered from a given position, where the cepstral parameters are used to effectively represent useful clean speech. Using the estimated frame sequence data, the GMM of the acoustic transfer function is created to deal with the influence of a room impulse response. Then, for each test dataset, we find a maximum-likelihood (ML) GMM from among the estimated GMMs corresponding to each position. The effectiveness of this method has been confirmed by talker localization experiments performed in a room environment.

1. Introduction

Many systems using microphone arrays have been tried in order to localize sound sources. Conventional techniques, such as MUSIC and CSP (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 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 and cost. 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]). In our previous work [11, 12], we proposed Hidden Markov Model (HMM) separation for reverberant speech recognition, where the observed (reverberant) speech is separated into the acoustic transfer function and the clean speech HMM. Using HMM separation, it is possible to estimate the acoustic transfer function using some adaptation data (only several words) uttered from a given position. For this reason, measurement of impulse responses is not required. Because the characteristics of the acoustic transfer function depend on each position, the obtained acoustic transfer function can be used to localize the talker.

In this paper, we will discuss a new talker localization method using only a single microphone. In our previous work [11] for reverberant speech recognition, HMM separation required texts of a user's utterances in order to estimate the acoustic transfer function. However, it is difficult to obtain texts of utterances for talker-localization estimation tasks. In this paper, the acoustic transfer function is estimated from observed (reverberant) speech using a clean speech model without having to rely on user utterance texts, where a Gaussian Mixture Model (GMM) is used to model clean speech features. This estimation is performed in the cepstral domain employing an approach based upon maximum likelihood (ML). This is possible because the cepstral parameters are an effective representation for retaining useful clean speech information. The results of our talker-localization experiments show the effectiveness of our method.

2. Estimation of the Acoustic Transfer Function

2.1. System Overview. Figure 1 shows the training process for the acoustic transfer function GMM. First, we record the reverberant speech data $O^{(\theta)}$ from each position θ in order to build the GMM of the acoustic transfer function for θ. Next, the frame sequence of the acoustic transfer function $\hat{H}^{(\theta)}$ is estimated from the reverberant speech $O^{(\theta)}$ (any utterance) using the clean speech acoustic model, where a GMM is used to model the clean speech feature:

$$\hat{H}^{(\theta)} = \arg\max_{H} \Pr(O^{(\theta)} \mid H, \lambda_S). \tag{1}$$

Here, λ_S denotes the set of GMM parameters for clean speech, while the suffix S represents the clean speech in the cepstral domain. The clean speech GMM enables us to estimate the acoustic transfer function from the observed speech without needing to have user utterance texts (i.e., text-independent acoustic transfer estimation). Using the estimated frame sequence data of the acoustic transfer function $\hat{H}^{(\theta)}$, the acoustic transfer function GMM for each position $\lambda_H^{(\theta)}$ is trained.

Figure 2 shows the talker localization process. For test data, the talker position $\hat{\theta}$ is estimated based on discrimination of the acoustic transfer function, where the GMMs of the acoustic transfer function are used. First, the frame sequence of the acoustic transfer function $\hat{H}$ is estimated from the test data (any utterance) using the clean speech acoustic model. Then, from among the GMMs corresponding to each position, we find a GMM having the ML in regard to $\hat{H}$:

$$\hat{\theta} = \arg\max_{\theta} \Pr(\hat{H} \mid \lambda_H^{(\theta)}), \tag{2}$$

where $\lambda_H^{(\theta)}$ denotes the estimated acoustic transfer function GMM for direction θ (location).

2.2. Cepstrum Representation of Reverberant Speech. The observed signal (reverberant speech), $o(t)$, in a room environment is generally considered as the convolution of clean speech and the acoustic transfer function:

$$o(t) = \sum_{l=0}^{L-1} s(t-l)h(l), \tag{3}$$

where $s(t)$ is a clean speech signal and $h(l)$ is an acoustic transfer function (room impulse response) from the sound source to the microphone. The length of the acoustic transfer function is L. The spectral analysis of the acoustic modeling is generally carried out using short-term windowing. If the length L is shorter than that of the window, the observed complex spectrum is generally represented by

$$O(\omega; n) = S(\omega; n) \cdot H(\omega; n). \tag{4}$$

However, since the length of the acoustic transfer function is greater than that of the window, the observed spectrum is approximately represented by $O(\omega; n) \approx S(\omega; n) \cdot H(\omega; n)$.

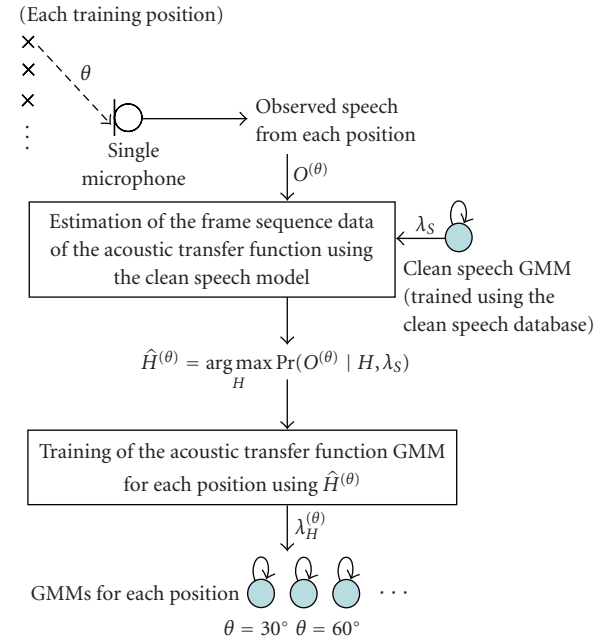

Figure 1: Training process for the acoustic transfer function GMM.

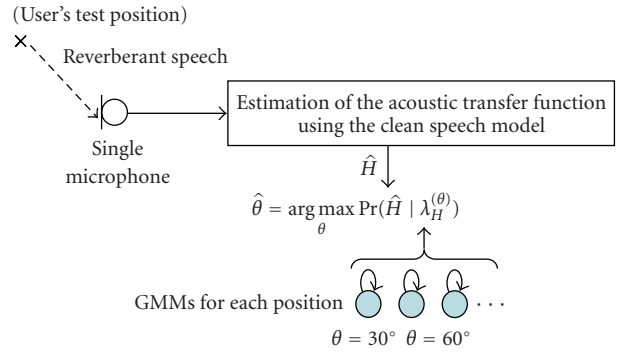

Figure 2: Estimation of talker localization based on discrimination of the acoustic transfer function.

Here $O(\omega; n)$, $S(\omega; n)$, and $H(\omega; n)$ are the short-term linear complex spectra in analysis window n. Applying the logarithm transform to the power spectrum, we get

$$\log|O(\omega; n)|^2 \approx \log|S(\omega; n)|^2 + \log|H(\omega; n)|^2. \tag{5}$$

In speech recognition, cepstral parameters are an effective representation when it comes to retaining useful speech information. Therefore, we use the cepstrum for acoustic modeling that is necessary to estimate the acoustic transfer function. The cepstrum of the observed signal is given by the inverse Fourier transform of the log spectrum:

$$O_{\text{cep}}(t; n) \approx S_{\text{cep}}(t; n) + H_{\text{cep}}(t; n), \tag{6}$$

where O_{cep}, S_{cep}, and H_{cep} are cepstra for the observed signal, clean speech signal, and acoustic transfer function, respectively. In this paper, we introduce a GMM of the acoustic transfer function to deal with the influence of a room impulse response.

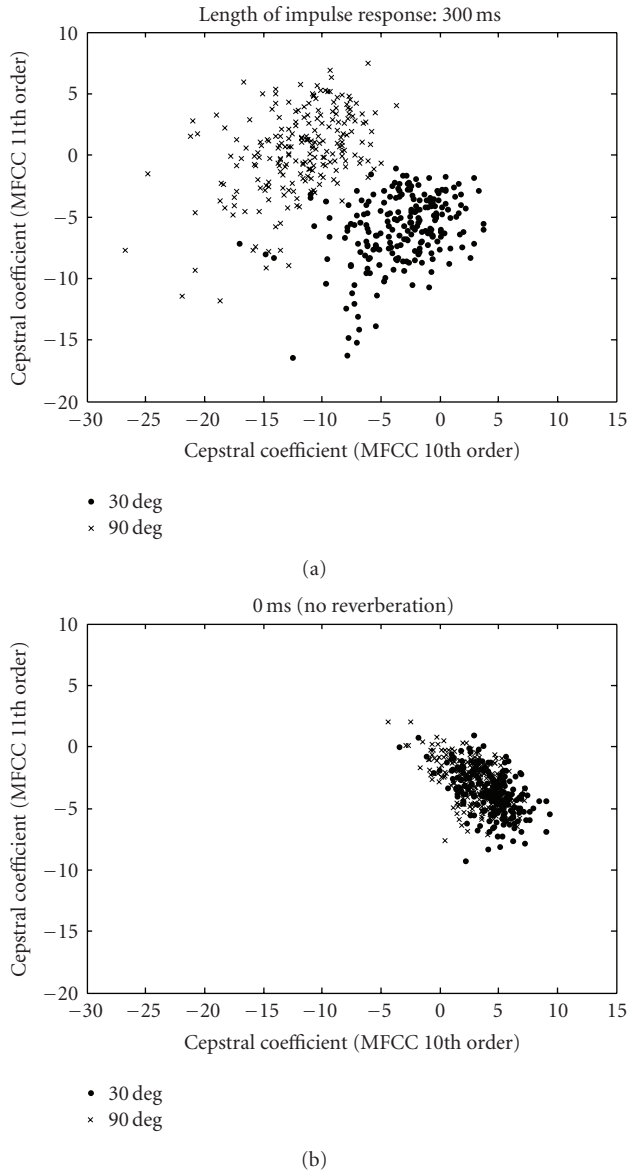

Figure 3: Difference between acoustic transfer functions obtained by subtraction of short-term-analysis-based speech features in the cepstrum domain.

2.3. Difference of Acoustic Transfer Functions. Figure 3 shows the mean values of the cepstrum, H'_{cep}, that were computed for each word using the following equations:

$$H_{\text{cep}}(t;n) \approx O_{\text{cep}}(t;n) - S_{\text{cep}}(t;n), \tag{7}$$

$$H'_{\text{cep}}(t) = \frac{1}{N}\sum_{n}^{N} H_{\text{cep}}(t;n), \tag{8}$$

where t is the cepstral index. Reverberant speech, O, was created using linear convolution of clean speech and impulse response. The impulse responses were taken from the RWCP sound scene database [13], where the loudspeaker was located at 30 and 90 degrees from the microphone. The lengths of the impulse responses are 300 and 0 milliseconds. The reverberant speech and clean speech were processed using a 32-millisecond Hamming window, and then for each frame, n, a set of 16 Mel-Frequency Cepstral Coefficients (MFCCs) was computed. The 10th and 11th cepstral coefficients for 216 words are plotted in Figure 3. As shown in this figure (300 milliseconds) a difference between the two acoustic transfer functions (30 and 90 degrees) appears in the cepstral domain. The difference shown will be useful for sound source localization estimation. On the other hand, in the case of the 0 millisecond impulse response, the influence of the microphone and the loudspeaker characteristics are a significant problem. Therefore, it is difficult to discriminate between each position for the 0 millisecond impulse response.

Also, this figure shows that the variability of the acoustic transfer function in the cepstral domain appears to be large for the reverberant speech. When the length of the impulse response is shorter than the analysis window used for the spectral analysis of speech, the acoustic transfer function obtained by subtraction of short-term-analysis-based speech features in the cepstrum domain comes to be constant over the whole utterance. However, as the length of the impulse response for the room reverberation becomes longer than the analysis window, the variability of the acoustic transfer function obtained by the short-term analysis will become large, with acoustic transfer function being approximately represented by (7). To compensate for this variability, a GMM is employed to model the acoustic transfer function.

3. Maximum-Likelihood-Based Parameter Estimation

This section presents a new method for estimating the GMM of the acoustic transfer function. The estimation is implemented by maximizing the likelihood of the training data from a user's position. In [14], an ML estimation method to decrease the acoustic mismatch for a telephone channel was described, and in [15] channel distortion and noise are simultaneously estimated using an expectation maximization (EM) method. In this paper, we introduce the utilization of the GMM of the acoustic transfer function based on the ML estimation approach to deal with a room impulse response.

The frame sequence of the acoustic transfer function in (6) is estimated in an ML manner by using the EM algorithm, which maximizes the likelihood of the observed speech:

$$\hat{H} = \arg\max_{H} \Pr(O \mid H, \lambda_S). \tag{9}$$

Here, λ_S denotes the set of clean speech GMM 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:

$$\begin{aligned} Q(\hat{H} \mid H) &= E[\log \Pr(O, c \mid \hat{H}, \lambda_S) \mid H, \lambda_S] \\ &= \sum_{c} \frac{\Pr(O, c \mid H, \lambda_S)}{\Pr(O \mid H, \lambda_S)} \cdot \log \Pr(O, c \mid \hat{H}, \lambda_S). \end{aligned} \tag{10}$$

Here c represents the unobserved mixture component labels corresponding to the observation sequence O.

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

$$\Pr(O, c \mid \hat{H}, \lambda_S) = \prod_{n^{(v)}} w_{c_{n^{(v)}}} \Pr(O_{n^{(v)}} \mid \hat{H}, \lambda_S), \tag{11}$$

where w is the mixture weight and $O_{n^{(v)}}$ is the cepstrum at the nth frame for the vth training data (observation data). Since we consider the acoustic transfer function as additive noise in the cepstral domain, the mean to mixture k in the model λ_O is derived by adding the acoustic transfer function. Therefore, (11) can be written as

$$\Pr(O, c \mid \hat{H}, \lambda_S) = \prod_{n^{(v)}} w_{c_{n^{(v)}}} \cdot N\left(O_{n^{(v)}}; \mu^{(S)}_{k_{n^{(v)}}} + \hat{H}_{n^{(v)}}, \Sigma^{(S)}_{k_{n^{(v)}}}\right), \tag{12}$$

where $N(O; \mu, \Sigma)$ denotes the multivariate Gaussian distribution. It is straightforward to derive that [16]

$$\begin{aligned} Q(\hat{H} \mid H) = {} & \sum_k \sum_{n^{(v)}} \Pr(O_{n^{(v)}}, c_{n^{(v)}} = k \mid \lambda_S) \log w_k \\ & + \sum_k \sum_{n^{(v)}} \Pr(O_{n^{(v)}}, c_{n^{(v)}} = k \mid \lambda_S) \\ & \cdot \log N\left(O_{n^{(v)}}; \mu^{(S)}_k + \hat{H}_{n^{(v)}}, \Sigma^{(S)}_k\right). \end{aligned} \tag{13}$$

Here $\mu^{(S)}_k$ and $\Sigma^{(S)}_k$ are the kth mean vector and the (diagonal) covariance matrix in the clean speech GMM, respectively. It is possible to train those parameters by using a clean speech database.

Next, we focus only on the term involving H:

$$\begin{aligned} & Q(\hat{H} \mid H) \\ & = \sum_k \sum_{n^{(v)}} \Pr\left(O_{n^{(v)}}, c_{n^{(v)}} = k \mid \lambda_S\right) \\ & \quad \cdot \log N\left(O_{n^{(v)}}; \mu^{(S)}_k + \hat{H}_{n^{(v)}}, \Sigma^{(S)}_k\right) \\ & = -\sum_k \sum_{n^{(v)}} \gamma_{k,n^{(v)}} \sum_{d=1}^{D} \left\{ \frac{1}{2} \log (2\pi)^D \sigma^{(S)^2}_{k,d} \right. \\ & \qquad \left. + \frac{\left(O_{n^{(v)},d} - \mu^{(S)}_{k,d} - \hat{H}_{n^{(v)},d}\right)^2}{2\sigma^{(S)^2}_{k,d}} \right\}, \\ & \gamma_{k,n^{(v)}} = \Pr(O_{n^{(v)}}, k \mid \lambda_S). \end{aligned} \tag{14}$$

Here D is the dimension of the observation vector O_n, and $\mu^{(S)}_{k,d}$ and $\sigma^{(S)^2}_{k,d}$ are the dth mean value and the dth diagonal variance value of the kth component in the clean speech GMM, respectively.

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

$$\hat{H}_{n^{(v)},d} = \frac{\sum_k \gamma_{k,n^{(v)}} \left(\left(O_{n^{(v)},d} - \mu^{(S)}_{k,d} \right) / \sigma^{(S)^2}_{k,d} \right)}{\sum_k \left(\gamma_{k,n^{(v)}} / \sigma^{(S)^2}_{k,d} \right)}. \tag{15}$$

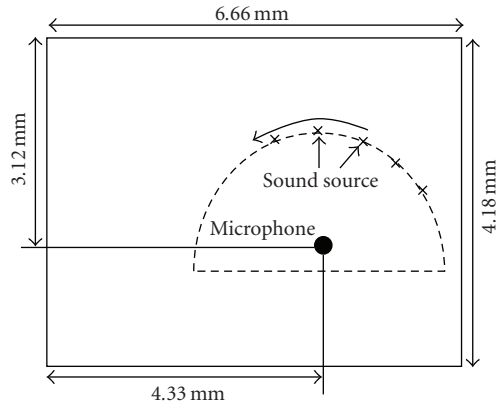

Figure 4: Experiment room environment for simulation.

After calculating the frame sequence data of the acoustic transfer function for all training data (several words), the GMM for the acoustic transfer function is created. The mth mean vector and covariance matrix in the acoustic transfer function GMM ($\lambda^{(\theta)}_H$) for the direction (location) θ can be represented using the term $\hat{H}_n$ as follows:

$$\begin{aligned} \mu^{(H)}_m &= \sum_v \sum_{n^{(v)}} \frac{\gamma_{m,n^{(v)}} \hat{H}_{n^{(v)}}}{\gamma_m}, \\ \Sigma^{(H)}_m &= \sum_v \sum_{n^{(v)}} \frac{\gamma_{m,n^{(v)}} \left(\hat{H}_{n^{(v)}} - \mu^{(H)}_m \right)^T \left(\hat{H}_{n^{(v)}} - \mu^{(H)}_m \right)}{\gamma_m}. \end{aligned} \tag{16}$$

Here $n^{(v)}$ denotes the frame number for vth training data.

Finally, using the estimated GMM of the acoustic transfer function, the estimation of talker localization is handled in an ML framework:

$$\hat{\theta} = \arg\max_{\theta} \Pr(\hat{H} \mid \lambda^{(\theta)}_H), \tag{17}$$

where $\lambda^{(\theta)}_H$ denotes the estimated GMM for θ direction (location), and a GMM having the maximum-likelihood is found for each test data from among the estimated GMMs corresponding to each position.

4. Experiments

4.1. 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 by a linear convolution of clean speech and impulse response. The impulse response was taken from the RWCP database in real acoustical environments [13]. The reverberation time was 300 milliseconds, and the distance to the microphone was about 2 meters. The size of the recording room was about 6.7 m × 4.2 m (width × depth). Figures 4 and 5 show the experimental room environment and the impulse response (90 degrees), respectively.

The speech signal was sampled at 12 kHz and windowed with a 32-millisecond Hamming window every 8 milliseconds. The experiment utilized the speech data of four males

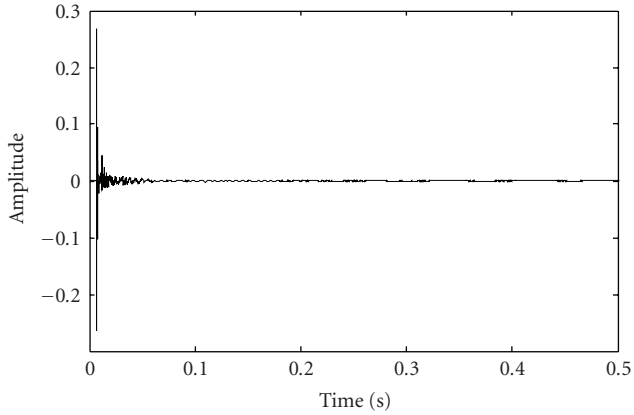

Figure 5: Impulse response (90 degrees, reverberation time: 300 milliseconds).

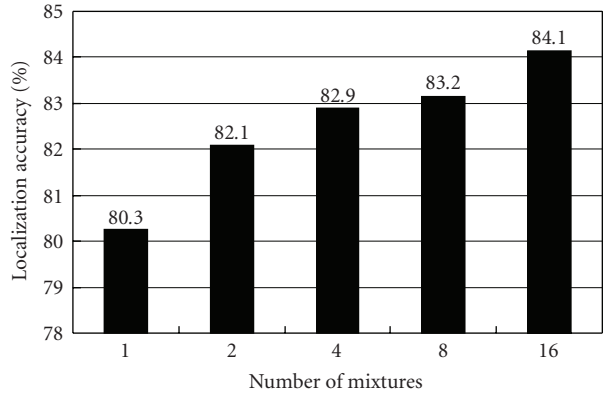

Figure 6: Effect of increasing the number of mixtures in modeling acoustic transfer function, here, 50 words are used for the estimation of the acoustic transfer function.

in the ATR Japanese speech database. The clean speech GMM (speaker-dependent model) was trained using 2620 words and has 64 Gaussian mixture components. The test data for one location consisted of 1000 words, and 16-order MFCCs were used as feature vectors. The total number of test data for one location was 1000 (words) $\times$ 4 (males). The number of training data for the acoustic transfer function GMM was 10 words and 50 words. The speech data for training the clean speech model, training the acoustic transfer function and testing were spoken by the same speakers but had different text utterances, respectively. The speaker's position for training and testing consisted of three positions (30, 90, and 130 degrees), five positions (10, 50, 90, 130, and 170 degrees), seven positions $(30, 50, 70, \ldots, 130$, and 150 degrees) and nine positions $(10, 30, 50, 70, \ldots, 150$, and 170 degrees). Then, for each set of test data, we found a GMM having the ML from among those GMMs corresponding to each position. These experiments were carried out for each speaker, and the localization accuracy was averaged by four talkers.

4.2. Performance in a Simulated Reverberant Environment. Figure 6 shows the localization accuracy in the three-position estimation task, where 50 words are used for the estimation of the acoustic transfer function. As can be seen from this figure, by increasing the number of Gaussian mixture components for the acoustic transfer function, the localization accuracy is improved. We can expect that the GMM for the acoustic transfer function is effective for carrying out localization estimation.

Figure 7 shows the results for a different number of training data, where the number of Gaussian mixture components for the acoustic transfer function is 16. The performance of the training using ten words may be a bit poor due to the lack of data for estimating the acoustic transfer function. Increasing the amount of training data (50 words) improves in the performance.

In the proposed method, the frame sequence of the acoustic transfer function is separated from the observed speech using (15), and the GMM of the acoustic transfer

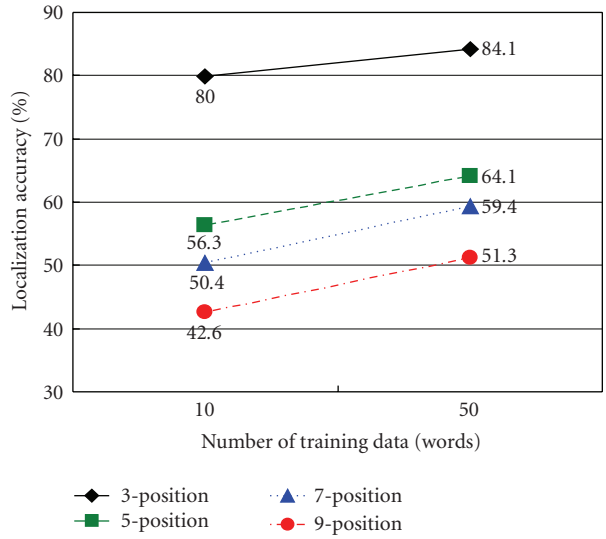

Figure 7: Comparison of the different number of training data.

function is trained by (16) using the separated sequence data. On the other hand, a simple way to carry out voice (talker) localization may be to use the GMM of the observed speech without the separation of the acoustic transfer function. The GMM of the observed speech can be derived in a similar way as in (16):

$$\mu_m^{(O)} = \sum_{\nu} \sum_{n^{(\nu)}} \frac{\gamma_{m,n^{(\nu)}} O_{n^{(\nu)}}}{\gamma_m},$$
$$\Sigma_m^{(O)} = \sum_{\nu} \sum_{n^{(\nu)}} \frac{\gamma_{m,n^{(\nu)}} \left(O_{n^{(\nu)}} - \mu_m^{(O)}\right)^T \left(O_{n^{(\nu)}} - \mu_m^{(O)}\right)}{\gamma_m}. \tag{18}$$

The GMM of the observed speech includes not only the acoustic transfer function but also clean speech, which is meaningless information for sound source localization. Figure 8 shows the comparison of four methods. The first method is our proposed method and the second is the

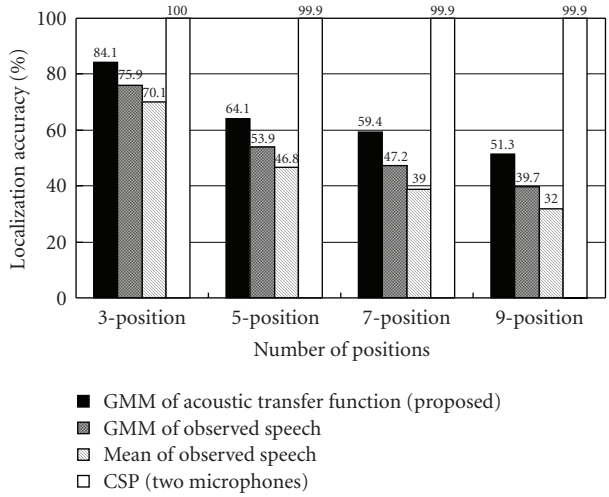

Figure 8: Performance comparison of the proposed method using GMM of the acoustic transfer function, a method using GMM of observed speech, that using the cepstral mean of observed speech, and CSP algorithm based on two microphones.

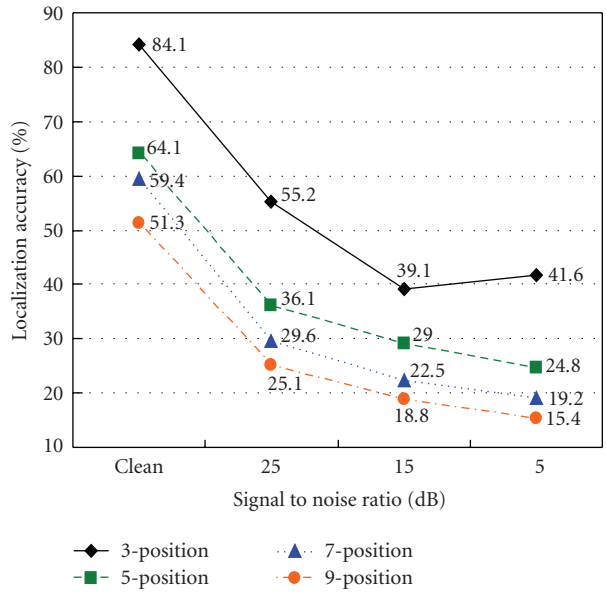

Figure 9: Localization accuracy for noisy environments.

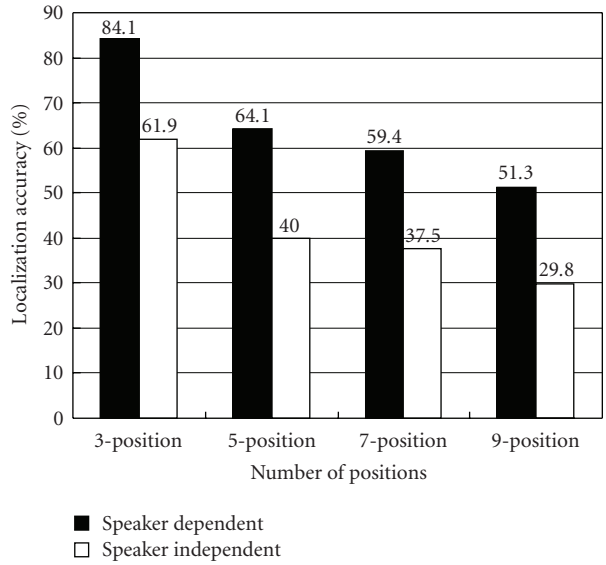

Figure 10: Comparison of performance using speaker-dependent/independent speech model (speaker-independent, 256 Gaussian mixture components; speaker-dependent, 64 Gaussian mixture components).

method using GMM of the observed speech without the separation of the acoustic transfer function. The third is a simpler method that uses the cepstral mean of the observed speech instead of GMM. (Then, the position that has the minimum distance from the learned cepstral mean to that of the test data is selected as the talker's position.) The fourth is a CSP (Cross-power Spectrum Phase) algorithm based on two microphones, where the CSP uses simultaneous phase information from microphone arrays to estimate the location of the arriving signal [2]. As shown in this figure, the use of the GMM of the observed speech had a higher accuracy than that of the mean of the observed speech, and, the use of the GMM of the acoustic transfer function results in a higher accuracy than that of GMM of the observed speech. The proposed method separates the acoustic transfer function from the short observed speech signal, so the GMM of the acoustic transfer function will not be affected greatly by the characteristics of the clean speech (phoneme). As it did with each test word, it is able to achieve good performance regardless of the content of the speech utterance, but the localization accuracy of the methods using just one microphone decreases as the number of training positions increases. On the other hand, the CSP algorithm based on two microphones has high accuracy even in the 9-position task. As the proposed method (single microphone only) uses the acoustic transfer function estimated from a user's utterance, the accuracy is low.

4.3. Performance in Simulated Noisy Reverberant Environments and Using a Speaker-Independent Speech Model. Figure 9 shows the localization accuracy for noisy environments. The observed speech data was simulated by adding pink noise to clean speech convoluted using the impulse response so that the signal to noise ratio (SNR) were 25 dB, 15 dB, and 5 dB. As shown in Figure 9, the localization accuracy at the SNR of 25 dB decreases about 30% in comparison to that in a noiseless environment. The localization accuracy decreases further as the SNR decreases.

Figure 10 shows the comparison of the performance between a speaker-dependent speech model and a speaker-independent speech model. For training a speaker-independent clean speech model and a speaker-independent acoustic transfer function model, the speech data spoken by four males in the ASJ Japanese speech database were used. Then, the clean speech GMM was trained using 160 sentences (40 sentences × 4 males) and it has 256 Gaussian

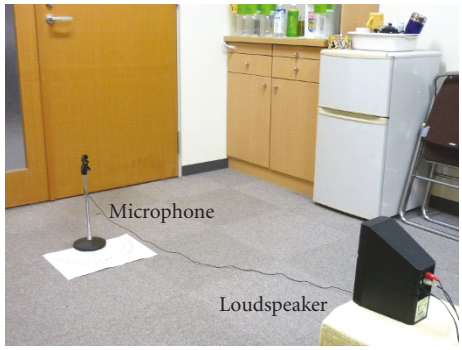

Figure 11: Experiment room environment.

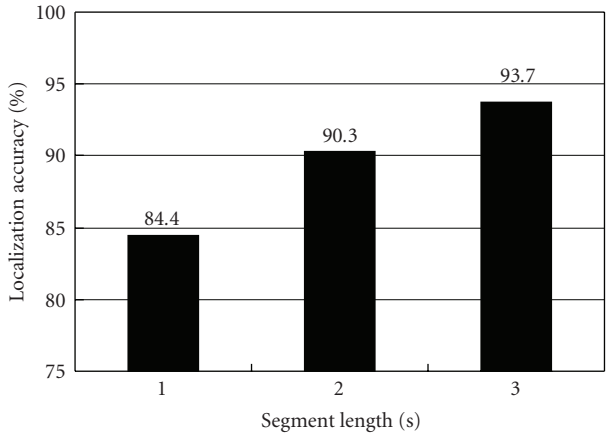

Figure 12: Comparison of performance using different test segment lengths.

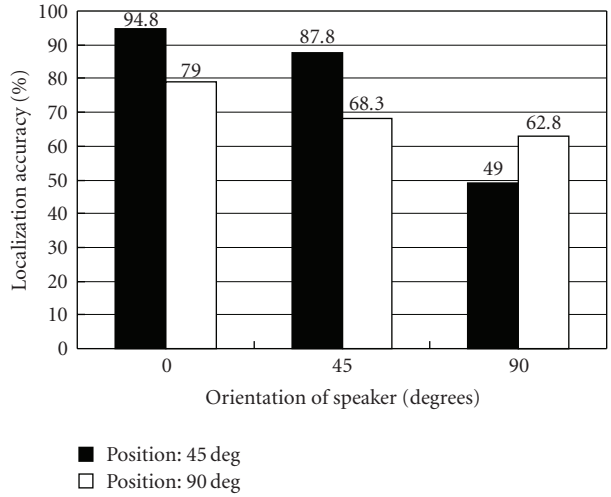

Figure 13: Effect of speaker orientation.

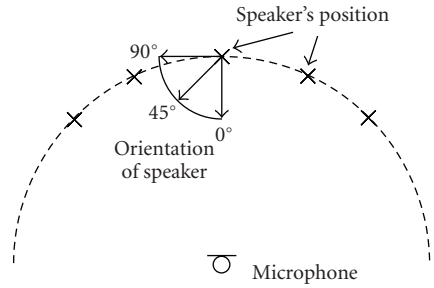

Figure 14: Speaker orientation.

mixture components. The acoustic transfer function for training locations was estimated by this clean speech model from 10 sentences for each male. The total number of training data for the acoustic transfer function GMM was 40 (10 sentences × 4 males) sentences. For training the speaker-dependent model and testing, the speech data spoken by four males in the ATR Japanese speech database were used in the same way as described in Section 4.1. The speech data for the test were provided by the same speakers used to train the speaker-dependent model, but different speakers were used to train the speaker-independent model. Both the speaker-dependent GMM and the speaker-independent GMM for the acoustic transfer function have 16 Gaussian mixture components. As shown in Figure 10, the localization accuracy of the speaker-independent speech model decreases about 20% in comparison to the speaker-dependent speech model.

4.4. Performance Using Speaker-Dependent Speech Model in a Real Environment. The proposed method, which uses a speaker-dependent speech model, was also evaluated in a real environment. The distance to the microphone was 1.5 m and the height of the microphone was about 0.45 m. The size of the recording room was about 5.5 m × 3.6 m × 2.7 m (width × depth × height). Figure 11 depicts the room environment of the experiment. The experiment used speech data, spoken by two males, in the ASJ Japanese speech database. The clean speech GMM (speaker-dependent model) was trained using 40 sentences and has 64 Gaussian mixture components. The test data for one location consisted of 200, 100, and 66 segments, where one segment has a time length of 1, 2, and 3 seconds, respectively. The number of training data for the acoustic transfer function was 10 sentences. The speech data for training the clean speech model, training the acoustic transfer function, and testing were spoken by the same speakers, but they had different text utterances, respectively. The experiments were carried out for each speaker and the localization accuracy of the two speakers was averaged.

Figure 12 shows the comparison of the performance using different test segment lengths. There were three speaker positions for training and testing (45, 90, and 135 degrees) and one loudspeaker (BOSE Mediamate II) was used for each position. As shown in this figure, the longer the length of the segment was, the more the localization accuracy increased, since the mean of estimated acoustic transfer function became stable. Figure 13 shows the effect when the orientation of the speaker changed from that of the

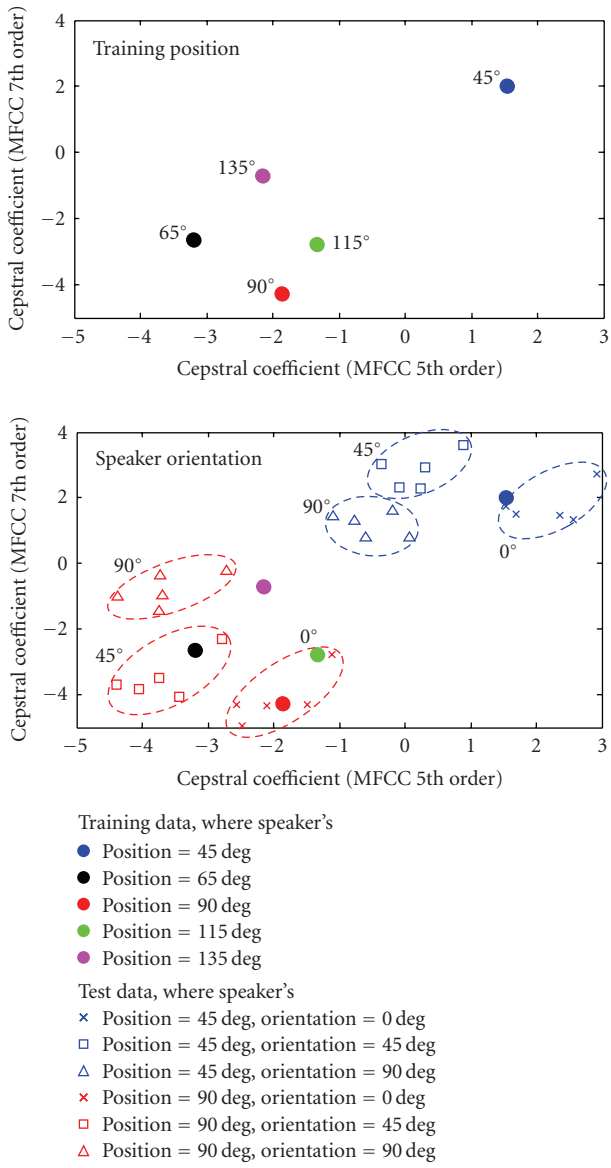

FIGURE 15: Mean acoustic transfer function values for five positions (top graph) and mean acoustic transfer function values for three speaker orientations (0 deg, 45 deg, and 90 deg) at a position of 45 deg and 90 deg (bottom graph).

speaker for training. There were five speaker positions for training (45, 65, 90, 115, and 135 degrees). There were two speaker positions for the test (45 and 90 degrees), and the orientation of the speaker changed to 0, 45, and 90 degrees, as shown in Figure 14. As shown in Figure 13, as the orientation of speaker changed, the localization accuracy decreased. Figure 15 shows the plot of acoustic transfer function estimated for each position and orientation of speaker. The plot of the training data is the mean value of all training data, and that for the test data is the mean value of test data per 40 seconds. As shown in Figure 15, as the orientation of the speaker changed from that for training, the estimated acoustic transfer functions were distributed over the distance away from the position of training data. As a result, these estimated acoustic transfer functions were not correctly recognized.

5. Conclusion

This paper has described a voice (talker) localization method using a single microphone. The sequence of the acoustic transfer function is estimated by maximizing the likelihood of training data uttered from a position, where the cepstral parameters are used to effectively represent useful clean speech information. The GMM of the acoustic transfer function based on the ML estimation approach is introduced to deal with a room impulse response. The experiment results in a room environment confirmed its effectiveness for location estimation tasks, but the proposed method requires the measurement of speech for each room environment in advance, and the localization accuracy decreases as the number of training positions increases. In addition, not only the position of speaker but also various factors (e.g., orientation of the speaker) affect the acoustic transfer function. Future work will include efforts to improve both localization estimation from more locations and estimation when the conditions other than speaker position change. We also hope to improve the localization accuracy in noisy environments and for speaker-independent speech models. Also, we will investigate a text-independent technique based on HMM in the modeling of the speech content.

References

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

[2] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using CSP analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, vol. 2, pp. 921–924, Atlanta, Ga, USA, May 1996.

[3] F. Asano, H. Asoh, and T. Matsui, "Sound source localization and separation in near field," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E83-A, no. 11, pp. 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 Transactions on Information and Systems*, vol. E89-D, no. 3, pp. 1050–1057, 2006.

[5] F. Keyrouz, Y. Naous, and K. Diepold, "A new method for binaural 3-D localization based on HRTFs," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, vol. 5, pp. 341–344, Toulouse, France, May 2006.

[6] M. Takimoto, T. Nishino, and K. Takeda, "Estimation of a talker and listener's positions in a car using binaural signals," in *Proceedings of the 4th Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan (ASA/ASJ '06)*, p. 3216, Honolulu, Hawaii, USA, November 2006, 3pSP33.

[7] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 2, pp. 817–820, Montreal, Canada, May 2004.

[8] B. Raj, M. V. S. Shashanka, and P. Smaragdis, "Latent dirichlet decomposition for single channel speaker separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, vol. 5, pp. 821–824, Toulouse, France, May 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," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 5, pp. 45–48, Hong Kong, April 2003.

[10] T. Nakatani, B.-H. Juang, K. Kinoshita, and M. Miyoshi, "Speech dereverberation based on probabilistic models of source and room acoustics," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, vol. 1, pp. 821–824, Toulouse, France, May 2006.

[11] T. Takiguchi, S. Nakamura, and K. Shikano, "HMM-separation-based speech recognition for a distant moving speaker," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 127–140, 2001.

[12] T. Takiguchi and M. Nishimura, "Acoustic model adaptation using first order prediction for reverberant speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 1, pp. 869–872, Montreal, Canada, May 2004.

[13] S. Nakamura, "Acoustic sound database collected for hands-free speech recognition and sound scene understanding," in *Proceedings of the International Workshop on Hands-Free Speech Communication (HSC '01)*, pp. 43–46, Kyoto, Japan, April 2001.

[14] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 3, pp. 190–202, 1996.

[15] T. Kristiansson, B. J. Frey, L. Deng, and A. Acero, "Joint estimation of noise and channel distortion in a generalized EM framework," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU '01)*, pp. 155–158, Trento, Italy, December 2001.

[16] B.-H. Juang, "Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains," *AT & T Technical Journal*, vol. 64, no. 6, pp. 1235–1249, 1985.