HMM-BASED SEPARATION OF ACOUSTIC TRANSFER FUNCTION FOR SINGLE-CHANNEL SOUND SOURCE LOCALIZATION

Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki

Graduate School of Engineering, Kobe University
1-1 Rokkodai, Nada-ku, Kobe, 657-8501 Japan

ABSTRACT

This paper presents a sound source (talker) localization method using only a single microphone, where a HMM (Hidden Markov Model) 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 (Gaussian Mixture Model) of the acoustic transfer function is created to deal with the influence of a room impulse response. Then, for each test data set, we find a maximum-likelihood 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.

Index Terms— single channel, talker localization, acoustic transfer function, maximum likelihood

1. INTRODUCTION

Many systems using microphone arrays have been tried in order to localize sound sources. Conventional techniques, such as MUSIC, CSP, and so on (e.g., [1, 2]), 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., [3, 4]). 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 actual car environments or 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., [5, 6]). In our previous work [7], we discussed a sound source localization method using only a single microphone. In that report, the acoustic transfer function was estimated from an observed (reverberant) speech using a clean speech model without texts of the user's utterance, where a GMM (Gaussian Mixture Model) 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. 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, where a HMM (Hidden Markov Model) of clean speech is used to estimate the acoustic transfer function from a user's position. Unlike GMM separation of our previous work, HMM separation requires texts of a user's utterances in order to estimate the acoustic transfer function. Therefore, the phoneme sequence of the observed (reverberant) signal is recognized first, and the recognition result is used as the text information to estimate the acoustic transfer function. This estimation is performed in the cepstral domain employing an approach based upon maximum likelihood. 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 phoneme sequence of the reverberant speech data is recognized by using each phoneme HMM of clean speech data. Using the recognition result, the phoneme HMMs are concatenated. And the frame sequence of the acoustic transfer function $\hat{H}^{(\theta)}$ is estimated from the reverberant speech $O^{(\theta)}$ using the concatenated HMM. 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 maximum-likelihood in regard to $\hat{H}$.

2.1.1. Cepstrum Representation of Reverberant Speech

The reverberant speech signal, $o(t)$, in a room environment is generally considered as the convolution of clean speech and acoustic transfer function. The spectral analysis of the acoustic modeling is generally carried out using short-term windowing. Therefore, the spectrum of the reverberant speech signal is approximately represented by $O(\omega;n) \approx S(\omega;n) \cdot H(\omega;n)$, where the length of the acoustic transfer function may be greater than that of the window. Here $O(\omega;n)$, $S(\omega;n)$, and $H(\omega;n)$ are the short-term linear spec-

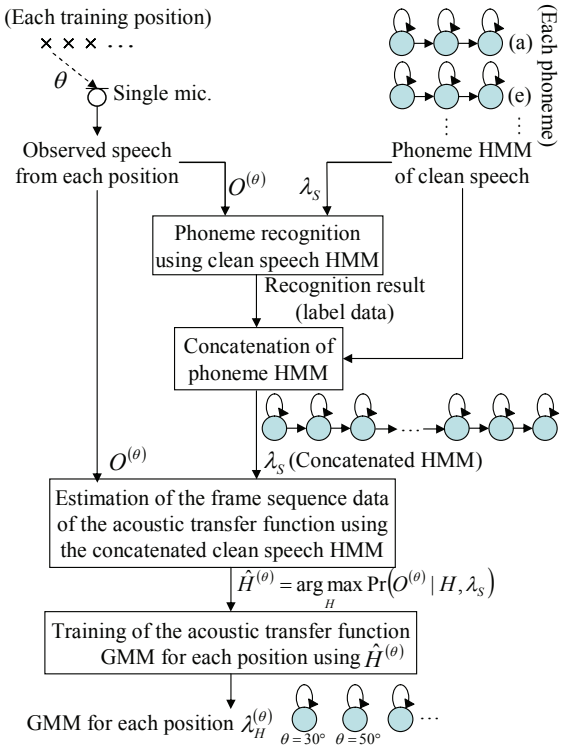

Fig. 1. Training process for the acoustic transfer function GMM

tra of the reverberant speech signal, clean speech signal, and the acoustic transfer function in the analysis window n, respectively.

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

$$O_{cep}(d;n) \approx S_{cep}(d;n) + H_{cep}(d;n) \tag{1}$$

where O_{cep}, S_{cep}, and H_{cep} are cepstra for the reverberant speech signal, clean speech signal, and acoustic transfer function, respectively. As shown in equation (1), if O and S are observed, H can be obtained by

$$H_{cep}(d;n) \approx O_{cep}(d;n) - S_{cep}(d;n). \tag{2}$$

However, S cannot be observed actually. Therefore, H is estimated by maximizing the likelihood (ML) of reverberant speech using clean-speech HMM.

2.2. Maximum-Likelihood-Based Parameter Estimation

This section presents a new method for estimating the GMM (Gaussian Mixture Model) of the acoustic transfer function. The estimation is implemented by maximizing the likelihood of the training data from a user's position. 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 (2) is estimated in an ML manner by using the expectation maximization (EM) algorithm, which maximizes the likelihood of the observed speech:

$$\hat{H} = \underset{H}{\operatorname{argmax}} \Pr(O|H, \lambda_S). \tag{3}$$

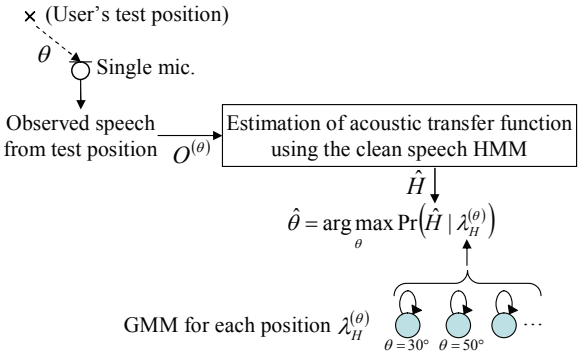

Fig. 2. Estimation of talker localization based on discrimination of the acoustic transfer function

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.

$$\begin{aligned} & Q(\hat{H}|H) \\ &= E[\log \Pr(O, p, b_p, c_p|\hat{H}, \lambda_S)|H, \lambda_S] \\ &= \textstyle\sum_p \sum_{b_p} \sum_{c_p} \frac{\Pr(O,p,b_p,c_p|H,\lambda_S)}{\Pr(O|H,\lambda_S)} \\ &\quad \cdot \log \Pr(O, p, b_p, c_p|\hat{H}, \lambda_S) \end{aligned} \tag{4}$$

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

$$\begin{aligned} & \Pr(O, p, b_p, c_p|\hat{H}, \lambda_S) \\ &= \textstyle\prod_n a_{b_p(n-1),b_p(n)} w_{c_p(n)} \Pr(O(n), p|\hat{H}, \lambda_S) \end{aligned} \tag{5}$$

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 to mixture k of state j in the model λ_O is derived by adding the acoustic transfer function. Therefore, (5) can be written as

$$\begin{aligned} & \Pr(O, p, b_p, c_p|\hat{H}, \lambda_S) \\ &= \textstyle\prod_n a_{b(n-1),b(n)} w_{b(n),c(n)} \\ &\quad \cdot N(O(n); \mu_{p,j,k}^{(S)} + \hat{H}(n), \Sigma_{p,j,k}^{(S)}) \end{aligned} \tag{6}$$

where $N(O; \mu, \Sigma)$ denotes the multivariate Gaussian distribution. It is straightforward to derive that [8]

$$\begin{aligned} & Q(\hat{H}|H) \\ &= \textstyle\sum_p \sum_i \sum_j \sum_n \\ &\quad \Pr(O(n), p, b_p(n) = j, b_p(n-1) = i|\lambda_S) \log a_{p,i,j} \\ &\quad + \textstyle\sum_p \sum_j \sum_k \sum_n \\ &\quad \Pr(O(n), p, b_p(n) = j, c_p(c) = k|\lambda_S) \log w_{p,j,k} \\ &\quad + \textstyle\sum_p \sum_j \sum_k \sum_n \\ &\quad \Pr(O(n), p, b_p(n) = j, c_p(n) = k|\lambda_S) \\ &\quad \cdot \log N(O(n); \mu_{p,j,k}^{(S)} + \hat{H}(n), \Sigma_{p,j,k}^{(S)}) \end{aligned} \tag{7}$$

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. It is possible to train those parameters by using a clean speech database.

Next, we focus only on the term involving H.

$$\begin{aligned} & Q(\hat{H}|H) \\ &= -\sum_p \sum_j \sum_k \sum_n \gamma_{p,j,k}(n) \\ &\quad -\sum_{d=1}^{D} \left\{ \frac{1}{2}\log(2\pi)^D \sigma^{(S)^2}_{p,j,k,d} \right. \\ &\quad \left. + \frac{(O(d;n)-\mu^{(S)}_{p,j,k,d}-\hat{H}(d;n))^2}{2\sigma^{(S)^2}_{p,j,k,d}} \right\} \end{aligned} \tag{8}$$

$$\gamma_{p,j,k}(n) = \Pr(O(n), p, j, k|\lambda_S) \tag{9}$$

Here D is the dimension of the observation vector O_n, and $\mu^{(S)}_{p,j,k,d}$ and $\sigma^{(S)^2}_{p,j,k,d}$ are the d-th mean value and the d-th diagonal variance value, respectively.

The maximization step (M-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

$$\hat{H}(d;n) = \frac{\sum_p \sum_j \sum_k \gamma_{p,j,k}(n) \frac{O(d;n)-\mu^{(S)}_{p,j,k,d}}{\sigma^{(S)^2}_{p,j,k,d}}}{\sum_p \sum_j \sum_k \frac{\gamma_{p,j,k}(n)}{\sigma^{(S)^2}_{p,j,k,d}}}. \tag{10}$$

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 m-th 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:

$$\mu^{(H)}_k = \sum_n \frac{\gamma_k(n)\hat{H}(n)}{\gamma_k} \tag{11}$$

$$\Sigma^{(H)}_m = \sum_n \frac{\gamma_k(n)(\hat{H}(n)-\mu^{(H)}_k)^T(\hat{H}(n)-\mu^{(H)}_k)}{\gamma_k} \tag{12}$$

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

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

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \Pr(\hat{H}|\lambda^{(\theta)}_H), \tag{13}$$

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.

3. EXPERIMENTS

3.1. Experimental Conditions

The new talker localization method was evaluated a simulated reverberant 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 [9]. The reverberation time was 300 msec, 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). Figure 3 shows the experimental room environment.

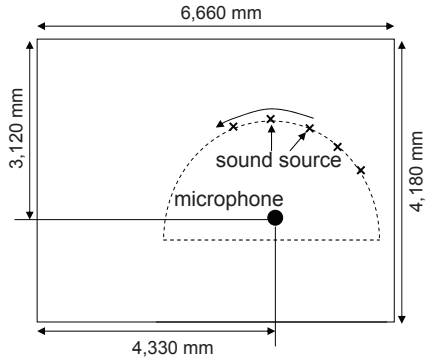

Fig. 3. Experiment room environment for simulation

The speech signal was sampled at 12 kHz and windowed with a 32-msec Hamming window every 8 msec. 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 and each phoneme HMM has 3 states and 32 Gaussian mixture components. The test data for one location consisted of 1,000 words, and 16-order MFCCs (Mel-Frequency Cepstral Coefficients) were used as feature vectors. The total number of test data for one location was 1,000 (words) × 5 (males). 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, ..., 130 and 150 degrees) and nine positions (10, 30, 50, 70, ..., 150, and 170 degrees). Then, for each set of test data, we found a GMM having the maximum-likelihood from among those GMMs corresponding to each position. These experiments were carried out for each speaker, and the localization accuracy was averaged by five talkers.

3.2. Experimental Results

The proposed method was compared with the other two method. One is the our previous method using GMM of the acoustic transfer function separated by the clean speech GMM. In this method, clean speech GMM was trained using the same clean speech data as that of the proposed method, and has 64 Gaussian mixture components. Another one is a simple way using the GMM of the observed speech without the separation of the acoustic transfer function. 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. Then, the GMM of the acoustic transfer function and the observed speech in each method was trained using 50 words and has 16 Gaussian mixture components.

As shown in Figure 4, the use of the GMM of the acoustic transfer function showed higher accuracies than that of the observed speech. This is because the GMM of the acoustic transfer function may not be affected greatly by the characteristics of the clean speech (phoneme). Also, the separation using the clean speech HMM showed higher accuracies than that using the clean speech GMM. Table 1 shows the mean square error (MSE) of the separated

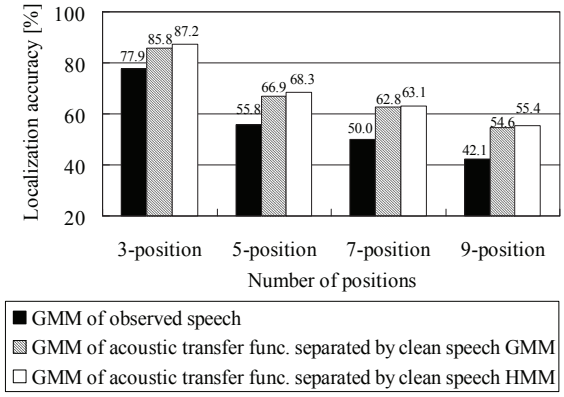

Fig. 4. Performance comparison of the separation using clean speech HMM, the separation using clean speech GMM, and without separation.

Table 1. Mean square error of the separated acoustic transfer function

	HMM	GMM
MSE	2096.14	2264.33

acoustic transfer function estimated by our proposed method and precious method, where the acoustic transfer function calculated by (2) using the true clean speech data is used as the ground truth. As shown in this table, the clean speech HMM can estimate the acoustic transfer function more correctly than clean speech GMM.

Also, each localization accuracy decreases as the number of training positions increases. Figure 5 and Figure 6 show the mean acoustic transfer function values for three and seven positions, respectively. The acoustic transfer functions are calculated by (2). As shown in these figures, when the number of position is three, the distribution of the acoustic transfer function for each position can be discriminated relatively-easily. However, when the number of position is seven, it is difficult to discriminate the distribution for each position. Therefore, when the the number of training positions increases, it is difficult to estimate the talker's position by this method.

4. CONCLUSION

This paper has described a voice (talker) localization method using a single microphone. The sequence of the acoustic transfer function is estimated by HMM (Hidden Markov Model) of clean speech. The experiment results in a room environment confirmed its effectiveness for location estimation tasks. But the localization accuracy decreases as the number of training positions increases. Therefore, we will research the feature vector retaining useful information to discriminate the acoustic transfer function for each position. 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 investigate the estimation when the conditions other than speaker position change.

5. REFERENCES

[1] D. Johnson and D. Dudgeon, *Array Signal Processing*, Prentice Hall, 1996.

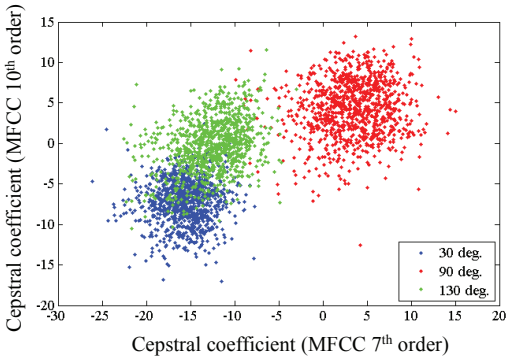

Fig. 5. Mean acoustic transfer function values for three positions.

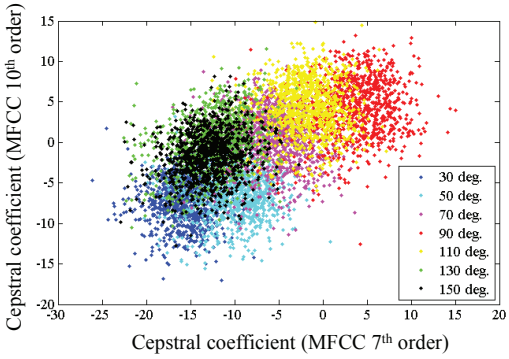

Fig. 6. Mean acoustic transfer function values for seven positions.

[2] M. Omologo and P. Svaizer, "Acoustic event localization in noisy and reverberant environment using csp analysis," in *Proc. ICASSP96*, 1996, pp. 921–924.

[3] F. Keyrouz, Y. Naous, and K. Diepold, "A new method for binaural 3-d localization based on hrtfs," in *Proc. ICASSP06*, 2006, pp. V–341–V–344.

[4] M. Takimoto, T. Nishino, and K. Takeda, "Estimation of a talker and listener's positions in a car using binaural signals," in *The Fourth Joint Meeting ASA and ASJ*, 2006, p. 3216.

[5] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proc. ICASSP04*, 2004, pp. 817–820.

[6] B. Raj, M. V. S. Shashanka, and P. Smaragdis, "Latent direchlet decomposition for single channel speaker separation," in *Proc. ICASSP06*, 2006, pp. 821–824.

[7] T. Takiguchi, Y. Sumida, R. Takashima, and Y. Ariki, "Single-channel talker localization based on discrimination of acoustic transfer functions," in *EURASIP Journal on Advances in Signal Processing*, 2009, vol. Volume 2009, p. 9pages.

[8] B.-H. Juang, "Maximum-likelihood estimation of mixture multivariate stochastic observations of markov chains," in *AT&T Tech. J.*, 1985, vol. Vol. 64, pp. 1235–1249.

[9] S. Nakamura, "Acoustic sound database collected for hands-free speech recognition and sound scene understanding," in *International Workshop on Hands-Free Speech Communication*, 2001, pp. 43–46.