ESTIMATION OF ROOM ACOUSTIC TRANSFER FUNCTION USING SPEECH MODEL

Tetsuya Takiguchi, Yuji Sumida, Yasuo Ariki

Department of Computer and System Engineering
Kobe University, Japan

ABSTRACT

This paper presents a sound source localization method using only a single microphone, where the GMM (Gaussian Mixture Model) of clean speech is introduced to estimate the acoustic transfer function from a user's position. The new method is able to estimate it without measuring impulse responses. 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 due to effectively represent useful clean speech. Using the estimated 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 data, we find a GMM having the maximum-likelihood from among the estimated GMMs corresponding to each position. Its effectiveness is confirmed by talker direction experiments in a room environment.

Index Terms— Direction of arrival estimation, Speech processing, Maximum likelihood estimation

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 signal arrival. 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.

Single-microphone source separation problem is one of the most challenging scenarios in the signal processing, and some techniques are described in literature, for example [3, 4, 5], where two-speaker separation or music-source separation techniques are introduced. In our previous work [6, 7], we proposed HMM (Hidden Markov Model) separation for estimating the HMM parameters of the acoustic transfer function on the basis of a maximum likelihood manner, where the observed (reverberant) speech is separated into the acoustic transfer function and the clean speech HMM. The HMM separation is able to estimate the acoustic transfer function using some adaptation data (only several sentences) uttered from a position. Therefore, measurement of impulse responses is not required. As the characteristic of the acoustic transfer function estimated by HMM separation depends on each position, the obtained acoustic transfer function will be useful for the talker localization.

In this paper, we will discuss a new talker localization method using only a single microphone. In our previous work [6], the proposed method required texts of user's utterance in order to estimate the acoustic transfer function. In this paper, the acoustic transfer function is estimated from observed (reverberant) speech using clean speech model without texts of user's utterance, where a GMM (Gaussian Mixture Model) with a single state only is used to model the feature of the clean speech. This estimation is performed in the cepstral domain employing a maximum likelihood based approach. This is possible because the cepstral parameters are an effective representation to retain useful clean speech information. The results of our talker-direction experiments show its effectiveness.

2. ESTIMATION OF THE ACOUSTIC TRANSFER FUNCTION

2.1. System Overview

First, we record the reverberant speech data (several sentences) from each position in order to build the GMM of the acoustic transfer function for each position. Next, the sequence data of the acoustic transfer function is estimated from the reverberant speech (any utterance) using the clean-speech acoustic model. Using the estimated sequence data of the acoustic transfer function, the GMM for each position is trained.

Fig. 1 shows the talker direction estimation using the estimated GMM of the acoustic transfer function. In order to estimate the talker direction, the sequence of the acoustic transfer function is estimated from test data (any utterance) using the clean-speech acoustic model. Then, we find a GMM having the maximum-likelihood from among the estimated GMMs corresponding to each position.

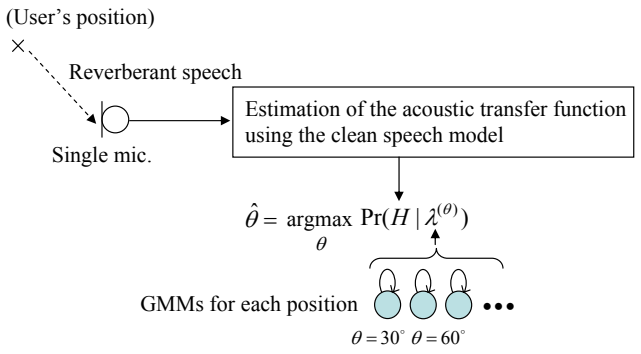

Fig. 1. Estimation of talker direction

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 acoustic transfer function:

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

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 spectrum is generally represented by

$$O(\omega; n) = S(\omega; n) \cdot H(\omega; n). \tag{2}$$

However, since the length of the acoustic transfer function is greater than that of the window, the observed spectrum is approximately represented by $O(\omega; n) \approx S(\omega; n) \cdot H(\omega; n)$. Here $O(\omega; n)$, $S(\omega; n)$, and $H(\omega; n)$ are the short-term linear spectra in the analysis window n. Applying the logarithm transform to the linear spectrum, we get

$$\log O(\omega; n) \approx \log S(\omega; n) + \log H(\omega; n). \tag{3}$$

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 observed signal is given by the inverse Fourier transform of the log spectrum.

$$O_{cep}(t; n) \approx S_{cep}(t; n) + H_{cep}(t; n) \tag{4}$$

where O_{cep}, S_{cep}, and H_{cep} are cepstra for the observed signal, clean speech signal, and acoustic transfer function. Since spectral analysis in acoustic modeling is based on short-term windowing, the multiplication of the short-term speech spectra and the acoustic transfer function is equivalent to periodic convolution in the time domain. However, the actual observed signal is the result of linear convolution. Therefore we cannot model the observed signal accurately. In this paper, we introduce a GMM (Gaussian Mixture Model) of the acoustic transfer function to deal with the influence of a room impulse response.

2.3. 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 training data (only several words) from a user's position. In [8], a maximum-likelihood (ML) estimation method is presented to decrease the acoustic mismatch for the telephone channel, where a single Gaussian mixture is used to model the channel mismatch. In this paper, we introduce the utilization of the GMM (Gaussian Mixture Model) of the acoustic transfer function based on the ML estimation approach to deal with a room impulse response.

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

$$\hat{H} = \underset{H}{\text{argmax}} \Pr(O|H, \lambda_S). \tag{5}$$

Here, λ denotes the set of 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.

$$\begin{aligned} Q(\hat{H}|H) &= E[\log \Pr(O, b, c|\hat{H}, \lambda_S)|H, \lambda_S] \\ &= \sum_b \sum_c \frac{\Pr(O, b, c|H, \lambda_S)}{\Pr(O|H, \lambda_S)} \cdot \log \Pr(O, b, c|\hat{H}, \lambda_S) \end{aligned} \tag{6}$$

Here b and c are the unobserved state sequence and the unobserved mixture component labels corresponding to the observation sequence O.

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

$$\Pr(O, b, c|\hat{H}, \lambda_S) = \prod_{n^{(v)}} a_{b_{n^{(v)}-1}, b_{n^{(v)}}} w_{b_{n^{(v)}}, c_{n^{(v)}}} \Pr(O_{n^{(v)}}|\hat{H}, \lambda_S) \tag{7}$$

where a is the transition probability, and w is the mixture weight. $O_{n^{(v)}}$ is the cepstrum at the n-th frame for the v-th 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, (7) can be written as

$$\Pr(O,b,c|\hat{H},\lambda_S) = \prod_{n^{(v)}} a_{b_{n^{(v)}-1},b_{n^{(v)}}} w_{b_{n^{(v)}},c_{n^{(v)}}} \cdot N(O_{n^{(v)}};\mu_{j,k_{n^{(v)}}}+\hat{H}_{n^{(v)}},\Sigma_{j,k_{n^{(v)}}}) \quad (8)$$

where $N(O;\mu,\Sigma)$ denotes the multivariate Gaussian distribution. It is straightforward to derive that [9]

$$\begin{aligned} &Q(\hat{H}|H) \\ &= \sum_i \sum_j \sum_{n^{(v)}} \Pr(O_{n^{(v)}}, b_{n^{(v)}}=j, b_{n^{(v)}-1}=i|\lambda_S)\log a_{i,j} \\ &\quad + \sum_j \sum_k \sum_{n^{(v)}} \Pr(O_{n^{(v)}}, b_{n^{(v)}}=j, c_{n^{(v)}}=k|\lambda_S)\log w_{j,k} \\ &\quad + \sum_j \sum_k \sum_{n^{(v)}} \Pr(O_{n^{(v)}}, b_{n^{(v)}}=j, c_{n^{(v)}}=k|\lambda_S) \\ &\qquad \cdot \log N(O_{n^{(v)}};\mu_{j,k}+\hat{H}_{n^{(v)}},\Sigma_{j,k}) \end{aligned} \quad (9)$$

Here we focus only on the term involving H.

$$\begin{aligned} &Q(\hat{H}|H) \\ &= \sum_j \sum_k \sum_{n^{(v)}} \Pr(O_{n^{(v)}}, b_{n^{(v)}}=j, c_{n^{(v)}}=k|\lambda_S) \\ &\quad \cdot \log N(O_{n^{(v)}};\mu_{j,k}+\hat{H}_{n^{(v)}},\Sigma_{j,k}) \\ &= -\sum_j \sum_k \sum_{n^{(v)}} \gamma_{j,k,n^{(v)}} \sum_{d=1}^{D} \left\{ \frac{1}{2}\log(2\pi)^D \sigma^2_{j,k,d} \right. \\ &\quad \left. + \frac{(O_{n^{(v)},d}-\mu_{j,k,d}-\hat{H}_{n^{(v)},d})^2}{2\sigma^2_{j,k,d}} \right\} \end{aligned} \quad (10)$$

$$\gamma_{j,k,n^{(v)}} = \Pr(O_{n^{(v)}}, j, k|\lambda_S) \quad (11)$$

Here D is the dimension of the adaptation vector O_n. 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

$$\hat{H}_{n^{(v)},d} = \frac{\sum_j \sum_k \gamma_{j,k,n^{(v)}} \dfrac{O_{n^{(v)},d}-\mu_{j,k,d}}{\sigma^2_{j,k,d}}}{\sum_j \sum_k \dfrac{\gamma_{j,k,n^{(v)}}}{\sigma^2_{j,k,d}}} \quad (12)$$

After the sequence data of the acoustic transfer function are calculated for all training data (several sentences), the GMM for the acoustic transfer function is created. The m-th mean vector and covariance matrix in the acoustic transfer function GMM ($\lambda_H^{(\theta)}$) for the direction θ can be represented by using the term $\hat{H}_n$ as follows:

$$\mu_m^{(H)} = \sum_v \sum_{n^{(v)}} \frac{\gamma_{m,n^{(v)}} \hat{H}_{n^{(v)}}}{\gamma_m} \quad (13)$$

$$\Sigma_m^{(H)} = \sum_v \sum_{n^{(v)}} \frac{\gamma_{m,n^{(v)}} (\hat{H}_{n^{(v)}}-\mu_m^{(H)})^T (\hat{H}_{n^{(v)}}-\mu_m^{(H)})}{\gamma_m} \quad (14)$$

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 the talker direction is handled in an ML framework:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \Pr(\hat{H}|\lambda_H^{(\theta)}), \quad (15)$$

where $\lambda_H^{(\theta)}$ denotes the estimated GMM for θ direction and we find a GMM having the maximum-likelihood 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 in a reverberant environment. 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 [10], where the target talker was located at 30, 90, and 130 degrees (test position). The reverberation time was 300 msec, and the distance to the microphone was about 2 m. The size of the recording room was about 6.7 m×4.2 m (width×depth).

The speech signal was sampled at 12 kHz and windowed with a 32-msec Hamming window every 8 msec. The clean speech GMM was trained by using 40 sentences spoken by one male in the ASJ Japanese speech database and has 64 Gaussian mixture components. The test data consisted of 100 (× 3 directions) sentences which are uttered from 30, 90, and 130 degrees, different from that used in the training, and 16-order MFCCs (Mel-Frequency Cepstral Coefficients) were used as feature vectors. The number of the training data for the acoustic transfer function GMM was one sentence, five sentences, and ten sentences which are uttered from 10, 30, 50, 70, ..., 150, and 170 degrees (nine directions). Therefore, nine GMMs are built, and then, for each test data, we find a GMM having the maximum-likelihood from among those GMMs corresponding to each position (nine directions).

3.2. Experimental Results

Figure 2 shows the direction accuracy in the nine-direction estimation task, where one sentence is 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 direction accuracy is improved. We can expect that the GMM for the acoustic

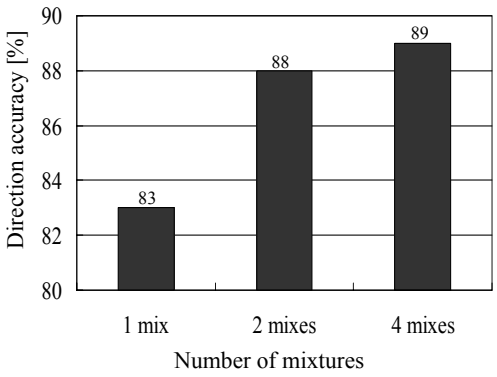

Fig. 2. Effect of increasing the number of mixtures in modelling acoustic transfer function. Here one sentence is used for the estimation of the acoustic transfer function.

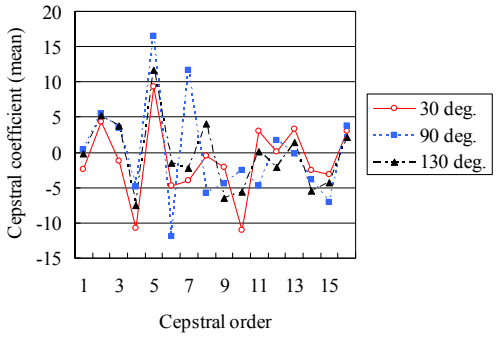

Fig. 3. Mean vectors of the different acoustic transfer functions

transfer function is effective for the direction estimation. Figure 3 shows the mean vectors (single mixture) of the different acoustic transfer functions from three training positions (directions). The differences shown will be useful for estimating of the sound source direction.

Figure 4 shows the results for the different number of training data. The performance of the training using one sentence is a little poor due to the lack of data for estimating the acoustic transfer function. Increasing the amount of training data improves in the performance and with about 5 sentences the acoustic transfer function appears to be estimated robustly.

Table 1 shows the direction accuracy for 0 msec and 300 msec (reverberation time) in the three-direction estimation task, where we find a GMM having the maximum-likelihood from among the GMMs for 30, 90, and 130 degrees. In the case of 300 msec, the direction accuracy is almost 100 %. But the direction accuracy for 0 msec decreases because the difference of the acoustic transfer function between each position becomes small. On the other hand, the CSP algorithm based on microphone arrays [2] has high accuracy (100 %) in the case of 0 msec. This is because the CSP uses simultaneous phase information from microphone arrays to estimate the direction of the signal arrival, and the proposed method (single microphone only) uses the acoustic transfer function to estimate the direction.

In the proposed method, the sequence of the acoustic transfer function is separated from the observed speech using (12), and using the separated sequence data, the GMM of the acoustic transfer function is trained by (13) and (14). On the other hand, a simple way for the voice (talker) localization may be the use of 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 (13)

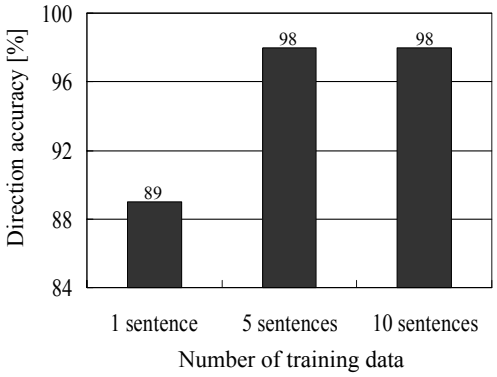

Fig. 4. Comparison of the different number of training data

and (14).

$$\mu_m^{(O)} = \sum_v \sum_{n^{(v)}} \frac{\gamma_{m,n^{(v)}} O_{n^{(v)}}}{\gamma_m} \tag{16}$$

$$\Sigma_m^{(O)} = \sum_v \sum_{n^{(v)}} \frac{\gamma_{m,n^{(v)}} (O_{n^{(v)}} - \mu_m^{(O)})^T (O_{n^{(v)}} - \mu_m^{(O)})}{\gamma_m} \tag{17}$$

The GMM of the observed speech includes not only the acoustic transfer function but also clean speech which is meaningless information for the sound source localization. As shown in Figure 5, the use of the GMM of the observed speech decreases the accuracy in comparison to that of the GMM of the acoustic transfer function, especially for the small training data. As the proposed method separates the acoustic transfer function from the observed speech, the use of the small training data only achieves good performance, and the GMM of the acoustic transfer function may not be much affected by the characteristics of the clean speech (phoneme) or the loud speaker characteristics.

Table 1. Direction accuracy for 0 msec reverberation time in the three-direction estimation task

Rev. time	1 mixture	2 mixtures	4 mixtures
0 msec	65.0 %	72.0 %	74.0 %
300 msec	99.7 %	99.7 %	100 %

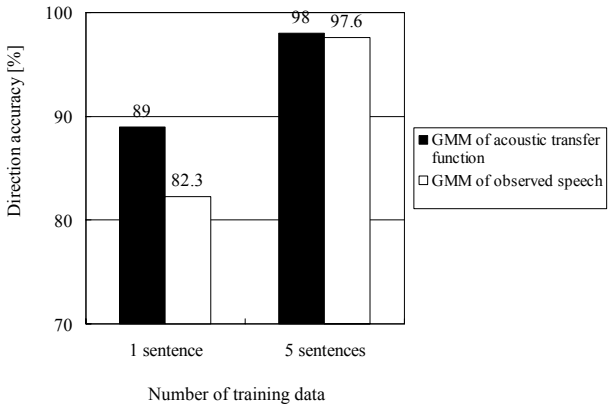

Fig. 5. Comparison of GMMs for the acoustic transfer function and the observed speech

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 maximizing the likelihood of training data (only several words) 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 the three-direction estimation task. Future work includes a direction estimation from among more directions, in noisy environments, and tests for speaker-independent speech model.

5. REFERENCES

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

[2] M. Omologo and P. Svaizer, "Acoustic Event Localization in Noisy and Reverberant Environment Using CSP Analysis," Proc. ICASSP96, pp. 921-924, 1996.

[3] T. Kristjansson, H. Attias and J. Hershey, "Single Microphone Source Separation Using High Resolution Signal Reconstruction," Proc. ICASSP04, pp. 817-820, 2004.

[4] B. Raj, M. V. S. Shashanka and P. Smaragdis, "Latent Direchlet Decomposition for Single Channel Speaker Separation," Proc. ICASSP06, pp. 821-824, 2006.

[5] G.-J. Jang. T.-W. Lee and Y.-H. Oh, "A Subspace Approach to Single Channel Signal Separation Using Maximum Likelihood Weighting Filters," Proc. ICASSP03, pp. 45-48, 2003.

[6] T. Takiguchi, S. Nakamura and K. Shikano, "HMM-separation-based speech recognition for a distant moving speaker," IEEE Transactions on Speech Audio Process., Vol. 9, No. 2, pp. 127-140, 2001.

[7] T. Takiguchi and M. Nishimura, "Acoustic Model Adaptation Using First Order Prediction for Reverberant Speech," Proc. ICASSP04, pp. 869-872, 2004.

[8] 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.

[9] B.-H. Juang, "Maximum-likelihood estimation of mixture multivariate stochastic observations of Markov chains," AT&T Tech. J., Vol. 64, No. 6, pp. 1235-1249, 1985.

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