SINGLE-CHANNEL MULTI-TALKER-LOCALIZATION BASED ON MAXIMUM LIKELIHOOD

Ryoichi Takashima, Tetsuya Takiguchi and Yasuo Ariki

Graduate School of Engineering, Kobe University, Kobe, Japan
takashima@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

ABSTRACT

This paper presents a sound source (talker) localization method using only a single microphone based upon maximum likelihood. In our previous work, we proposed GMM (Gaussian Mixture Model) separation for estimation of the sound source direction, where the observed (reverberant) speech is separated into the acoustic transfer function and the clean speech GMM, and showed its effectiveness for the single-talker localization task. In this paper, we discuss a multi-talker localization method using GMM separation and model composition. Model composition is used to represent speech signals observed in a reverberant environment corresponding to every conceivable combination of positions of the sound sources, where composite models are obtained through composition of talker's speech model and acoustic transfer functions estimated using GMM separation. For each test data set, we find a maximum-likelihood model from among the composite models corresponding to each combination of talkers' positions. The effectiveness of this method has been confirmed by two-talker localization experiments performed in a room environment.

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

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 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. 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 experiment results of single-talker-localization showed its effectiveness.

In this paper, we discuss a new single-channel multi-talker-localization method based upon maximum likelihood using model composition. For each talker, the models of clean speech and acoustic transfer functions estimated using GMM separation [7] are trained using GMM and SGM (Single Gaussian Model), respectively. Then, for each combination of talkers' position, the composite model of speech observed in a reverberant environment is obtained by composition of these GMMs and SGMs.

2. MODEL OF SPEECH OBSERVED IN A REVERBERANT ENVIRONMENT

2.1. System Overview

First, we record the reverberant speech data (several sentences) uttered by one people from each position in order to build the SGM 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 SGM for each position is trained. The SGM of the acoustic transfer function is trained for each talker because the estimated sequence data of the acoustic transfer function may be influenced by the phoneme sequence of clean speech of each talker.

Fig. 1 shows the composite model of observation speech, where the number of talkers is two. The composite GMM of the observed speech for each combination of talkers' position is obtained by composition of each talker's acoustic model

978-1-4244-2710-9/09/$25.00

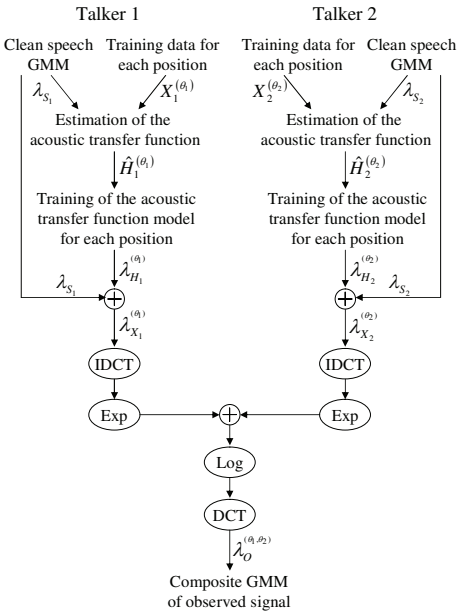

Fig. 1. Composite model of the observed speech

for clean speech and the acoustic transfer function. Then, for each test data set, we find a maximum-likelihood model from among the composite models corresponding to each combination of talkers' positions.

2.2. Estimation of the Acoustic Transfer Function

In our previous work, we proposed the method to estimate the acoustic transfer function from the reverberant speech (any utterance) using the clean-speech acoustic model, where a GMM is used to model the feature of the clean speech. The clean speech GMM enables us to estimate the acoustic transfer function from the observed speech without texts of user's utterance (text-independent estimation).

2.2.1. Cepstrum Representation of Reverberant Speech

The reverberant speech signal, $x(t)$, 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 $X(\omega;n) \approx S(\omega;n) \cdot H(\omega;n)$, where the length of the acoustic transfer function may be greater than that of the window. Here $X(\omega;n)$, $S(\omega;n)$, and $H(\omega;n)$ are the short-term linear spectra 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.

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

where X_{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 X and S are observed, H can be obtained by

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

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

2.2.2. Maximum-Likelihood-Based Parameter Estimation

The sequence of the acoustic transfer function in (2) is estimated in an ML manner using the expectation maximization (EM) algorithm, which maximizes the likelihood of the reverberant speech:

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

Here, λ denotes the set of GMM parameters of the clean speech, while the suffix S represents the clean speech in the cepstral domain. The GMM of clean speech consists of a mixture of Gaussian distributions.

$$\lambda_S = \{w_k, N(\mu_k^{(S)}, {\sigma_k^{(S)}}^2)\}, \qquad \textstyle\sum_k w_k = 1 \tag{4}$$

where w_k, μ_k and σ_k^2 are the weight coefficient, mean vector and variance vector (diagonal covariance matrix) of the k-th mixture component, respectively. Those parameters are estimated by EM (Expectation-Maximization) algorithm using a clean speech database.

The estimation of the acoustic transfer function in each frame is performed in a maximum likelihood fashion using the EM algorithm. The EM algorithm is a two-step iterative procedure. In the first step, called the expectation step, the auxiliary function Q is computed. Then, Q is defined as follows [7]:

$$\begin{aligned} Q(\hat{H}|H) \;=\; & -\textstyle\sum_k \sum_n \gamma_k(n) \sum_{d=1}^{D} \Big\{ \frac{1}{2}\log(2\pi)^D {\sigma_{k,d}^{(S)}}^2 \\ & + \frac{(X(d;n)-\mu_{k,d}^{(S)}-\hat{H}(d;n))^2}{2{\sigma_{k,d}^{(S)}}^2} \Big\} \end{aligned} \tag{5}$$

$$\gamma_k(n) = \Pr(X(n), k|\lambda_S) \tag{6}$$

Here, $X(n)$ is the cepstrum at the n-th frame for reverberant speech data. D is the dimension of the $X(n)$, and $\mu_{k,d}^{(S)}$ and ${\sigma_{k,d}^{(S)}}^2$ are the d-th mean value and the d-th diagonal variance value of the k-th component in the clean speech GMM, 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

$$\hat{H}(d;n) = \frac{\sum_k \gamma_k(n)\frac{X(d;n)-\mu_{k,d}^{(S)}}{\sigma_{k,d}^{(S)^2}}}{\sum_k \frac{\gamma_k(n)}{\sigma_{k,d}^{(S)^2}}}. \quad (7)$$

2.3. Model Composition

Using the estimated sequence data of the acoustic transfer function, the SGM for each position is trained. Then, using the obtained acoustic model parameter of the acoustic transfer function $\lambda_{H_i}^{\theta_i} = \left\{\mu^{(H_i^{\theta_i})}, \Sigma^{(H_i^{\theta_i})}\right\}$ and that of clean speech λ_{S_i}, the composite model of observed signal λ_O^{Θ} ($\Theta = \{\theta_1, \ldots, \theta_M\}$) for each combination of talkers' position is obtained [8].

The spectrum of the observed signal is expressed as

$$O(\omega;n) = \sum_{i=1}^{M} X_i^{\theta_i}(\omega;n) \quad (8)$$

$$X_i^{\theta_i}(\omega;n) \approx S_i(\omega;n)\cdot H_i^{\theta_i}(\omega;n). \quad (9)$$

where $X_i^{\theta_i}(\omega;n)$ is the spectrum of reverberant speech signal uttered by each talker i ($i = 1\cdots M$) from each talker's position θ_i, $H_i^{\theta_i}$ is the acoustic transfer function from the position, θ_i, and S_i is the clean speech signal of each talker i. First, using the acoustic model parameters of S_i and $H_i^{\theta_i}$ trained in the cepstral domain, the acoustic model of the reverberant speech signal $\lambda_{X_i}^{\theta_i}$ is obtained. As shown in equation (1), the mean vector and covariance matrix of the reverberant speech model are obtained as

$$\mu_{\rm cep}^{(X_i^{\theta_i})} = \mu_{\rm cep}^{(S_i)} + \mu_{\rm cep}^{(H_i^{\theta_i})},\ \Sigma_{\rm cep}^{(X_i^{\theta_i})} = \Sigma_{\rm cep}^{(S_i)} + \Sigma_{\rm cep}^{(H_i^{\theta_i})}. \quad (10)$$

Next, using the reverberant speech model, $\lambda_{X_i}^{\theta_i}$, obtained by equation (10) for each talker, the model of the observed signal λ_O^{Θ} is obtained. As shown in equation (8), the observed signal is represented by addition of the reverberant signal in the spectral domain. So, $\lambda_{X_i}^{\theta_i}$ needs to be transformed from the cepstral domain to the linear-spectral domain. The acoustic model parameter in the log-spectral domain is obtained by computing the inverse cosine transform of each Gaussian probability density function (PDF) of the GMM.

$$\mu_{\rm log}^{(X_i^{\theta_i})} = \Gamma^{-1}\mu_{\rm cep}^{(X_i^{\theta_i})},\ \Sigma_{\rm log}^{(X_i^{\theta_i})} = \Gamma^{-1}\Sigma_{\rm cep}^{(X_i^{\theta_i})}(\Gamma^{-1})^T \quad (11)$$

Here, Γ is a cosine transform matrix, $\mu_{\rm log}^{(X_i^{\theta_i})}$ and $\Sigma_{\rm log}^{(X_i^{\theta_i})}$ are the mean vector and covariance matrix of a Gaussian PDF in the log-power spectral domain, respectively. Then, the model parameter of observed signal in the liner-spectral domain is obtained by computing the exponential transform of reverberant signal model and adding these model parameter.

$$\mu_{{\rm lin},p}^{(X_i^{\theta_i})} = \exp\left\{\mu_{{\rm log},p}^{(X_i^{\theta_i})} + \sigma_{{\rm log},pp}^{(X_i^{\theta_i})^2}/2\right\}$$

$$\sigma_{{\rm lin},pq}^{(X_i^{\theta_i})^2} = \mu_{{\rm lin},p}^{(X_i^{\theta_i})}\cdot\mu_{{\rm lin},q}^{(X_i^{\theta_i})}\cdot\exp\left\{\sigma_{{\rm log},pq}^{(X_i^{\theta_i})^2} - 1\right\} \quad (12)$$

$$\mu_{\rm lin}^{(O^{\Theta})} = \sum_{i=1}^{M}\mu_{\rm lin}^{(X_i^{\theta_i})},\quad \Sigma_{\rm lin}^{(O^{\Theta})} = \sum_{i=1}^{M}\Sigma_{\rm lin}^{(X_i^{\theta_i})} \quad (13)$$

Here, $\mu_{{\rm lin},p}^{(X_i^{\theta_i})}$ and $\sigma_{{\rm lin},pq}^{(X_i^{\theta_i})^2}$ are the p-th mean and (p,q) element of the covariance matrix in the linear-spectral domain, respectively. Finally, the model parameter of observed signal in the cepstral domain is obtained by computing the log transform and cosine transform.

$$\sigma_{{\rm log},pq}^{(O^{\Theta})^2} = \log\left\{\frac{\sigma_{{\rm lin},pq}^{(O^{\Theta})^2}}{\mu_{{\rm lin},p}^{(O^{\Theta})}\cdot\mu_{{\rm lin},q}^{(O^{\Theta})}} + 1\right\}$$

$$\mu_{{\rm log},p}^{(O^{\Theta})} = \log\mu_{{\rm lin},p}^{(O^{\Theta})} - \sigma_{{\rm log},pp}^{(O^{\Theta})^2}/2 \quad (14)$$

$$\mu_{\rm cep}^{(O^{\Theta})} = \Gamma\mu_{\rm log}^{(O^{\Theta})},\ \Sigma_{\rm cep}^{(O^{\Theta})} = \Gamma\Sigma_{\rm log}^{(O^{\Theta})}\Gamma^T \quad (15)$$

2.4. Maximum-Likelihood-Based Localization

The acoustic model of observed signal is computed for all combination of talkers and those positions. Then, using the composite GMM of the observed signal, the estimation of talker localization is handled in an ML framework:

$$\hat{\Theta} = \underset{\Theta}{\rm argmax}\,\Pr(O|\lambda_{O^{\Theta}}) \quad (16)$$

where $\lambda_{O^{\Theta}}$ denotes the composite GMM for the combination of the direction (location) Θ ($\Theta = \{\theta_1, \ldots, \theta_M\}$), and a GMM having the maximum-likelihood is found for each test data from among the composite GMMs corresponding to each combination of talkers' positions.

3. EXPERIMENT

3.1. Experiment condition

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 [9], 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 talkers were one male and one female, and the number of position's combination was 9.

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 using 40 sentences spoken by one male in the ASJ Japanese speech database and has 64 Gaussian mixture components. Then 16-order MFCCs (Mel Frequency Cepstral Coefficients) were used as feature vectors.

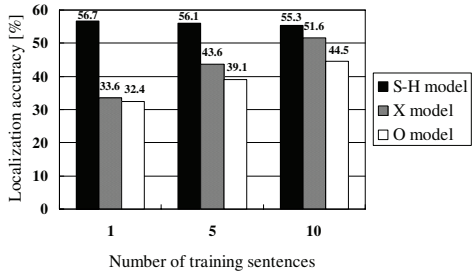

Fig. 2. Performance comparison of S-H model, X model and O model, where the both talkers' positions were estimated correctly.

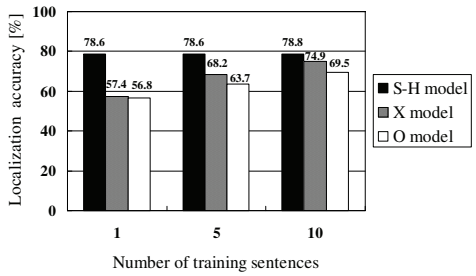

Fig. 3. Performance comparison of S-H model, X model and O model, where at least one talker's position was estimated correctly.

3.2. Experiment results

Fig. 2 and Fig. 3 show the comparison of three methods. The first method is our proposed method (S-H model), and the second (X model) is the method that $\lambda_{X_i}^{\theta_i}$ is trained using reverberant speech $X_i^{\theta_i}$ without separating into the acoustic transfer function and the clean speech GMM. The third (O model) is a simple method that λ_{O^Θ} is trained using observed speech signal directly, and this method needs the speech data uttered by all talkers at the same time for every combination of talkers' position instead of model composition. The number of the training data for the observed speech GMM was one sentence, five sentences, and ten sentences. The number of Gaussian mixture components for X model and O model was selected as the best accuracy was obtained in each experimental condition. Fig. 2 shows the performance of three methods, where the both talkers' positions were estimated correctly, and Fig. 3 shows the performance, where at least one talker's position was estimated correctly.

As shown in these figures, the localization accuracy of X model and O model decreases as the number of training data decreases. On the other hand, the localization accuracy of our proposed method is not so different even if the number of training data is changed. Therefore, once the clean speech model is trained, our proposed method does not need so many training data for the observed speech model.

4. CONCLUSION

This paper has described a new single-channel multi-talker-localization method based upon maximum likelihood using model composition. For each talker, the models of clean speech and acoustic transfer functions estimated using GMM separation are trained using GMM and SGM (Single Gaussian Model), respectively. For each combination of talkers' position, the composite model of observed speech is obtained by composition of these GMMs and SGMs. The experiment results in a room environment confirmed that the proposed method can estimate the localization by a few sentences of training data for each position. However, because the observed speech is influenced by the phoneme sequence of clean speech, proposed method cannot obtain so high localization accuracy. So, in future work, we will research to suppress the influence of the phoneme sequence of clean speech from observed speech signal uttered by multi-talker. Also, we will investigate for more talkers and more positions.

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] F. Keyrouz, Y. Naous, and K. Diepold, "A New Method for Binaural 3-D Localization Based on HRTFs," Proc. ICASSP06, pp. V-341-V-344, 2006.

[4] M. Takimoto, T. Nishino, and K. Takeda, "Estimation of a talker and listener's positions in a car using binaural signals," The Fourth Joint Meeting ASA and ASJ, 3pSP33, p. 3216, 2006.

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

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

[7] T. Takiguchi, Y. Sumida, and Y. Ariki, "Estimation of Room Acoustic Transfer Function Using Speech Model," IEEE Statistical Signal Processing Workshop, pp. 336-340, 2007.

[8] T. Takiguchi, M. Nishimura, and Y Ariki, "Acoustic Model Adaptation Using First-Order Linear Prediction for Reverberant Speech, " IEICE Trans. INF. and SYST., vol. E89-D, 908-914, 2006.

[9] 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.