Single-channel Head Orientation Estimation Based on Discrimination of Acoustic Transfer Function

Ryoichi Takashima, Tetsuya Takiguchi and Yasuo Ariki

Graduate School of System Informatics, Kobe University, Japan

takashima@me.cs.scitec.kobe-u.ac.jp, takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

Abstract

This paper presents a talker's head orientation estimation method using only a single microphone, where phoneme HMMs (Hidden Markov Models) of clean speech are introduced to separate the acoustic transfer function at the user's position and head orientation. The frame sequence of the acoustic transfer function is estimated by maximizing the likelihood of training data uttered from a given position with a given head orientation. Using the separated frame sequence data, the user's position and the head orientation are trained by Support Vector Machine (SVM) in advance. Then, for each test utterance, the frame sequence of the acoustic transfer function is separated based on the maximum likelihood estimation using the label sequence obtained from the phoneme recognition, and the user's position and head orientation are estimated by discriminating the separated acoustic transfer function using SVM. The effectiveness of this method has been confirmed by talker localization and head orientation estimation experiments performed in a real environment.

Index Terms: single channel, talker localization, head orientation, acoustic transfer function

1. Introduction

For human-human or human-computer interaction, the talker's head orientation is an important cue that determines not only who is talking but also who he/she is talking to. This who-talks-to-whom information can be helpful especially in multi-user conversation scenarios, such as a meeting system and the discrimination of system requests or users' conversations.

Many systems have been tried in order to localize sound sources. On the other hand, interest in the head orientation estimation from speech signals is relatively recent, and some approaches have been described [1, 2, 3, 4]. These methods use a network of microphone arrays in order to estimate the talker's head orientation. The approach described in [1] is based on the SRP-PHAT algorithm, which is often used for talker localization. In that paper, they modify the SRP-PHAT function by combining it with the weight function depending on the talker's head orientation. Other approaches focus on the radiation pattern of the magnitude for each head orientation of the talker [2, 3]. A method has also been proposed using the Direction-of-Arrival (DOA) histogram made from the DOA estimation results [4]. However, microphone array network systems need to be set along the walls of a given room so that sub-microphone arrays surround the user, and these systems may not be suitable in some cases due to their size and cost. Therefore, single-channel techniques are of interest, especially in small-device-based scenarios.

In our previous work [5], we discussed a sound source localization method using only a single microphone. In that report, the acoustic transfer function was estimated from observed (reverberant) speech using a clean speech model without texts of the user's utterances, and an HMM was used to model the features of the clean speech. 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. 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. Using the estimated frame sequence data, the user's position is trained, and for each test utterance, the user's position is estimated by discriminating the separated acoustic transfer function in the same way.

However, the impulse response may depend not only on the talker's position but also the head orientation. Therefore, in this paper, we will discuss a single-channel head orientation estimation method based on the discrimination of the acoustic transfer function. The proposed method trains the given pair of the talker's position and the head orientation, while our previous work trains only the talker's position using the estimated frame sequence of the acoustic transfer function. Compared with the other published works, this method requires a training process using a few observed speech utterances in advance. However, our proposed method is able to set a microphone anywhere in the given room. The effectiveness of this method has been confirmed by talker localization and head orientation estimation experiments performed in a real room environment.

2. Proposed Method

2.1. System Overview

Figure 1 shows the system overview. First, we record the reverberant speech data $O_{train}^{(\phi,\theta)}$ uttered from each position ϕ with each head orientation θ in order to train the acoustic transfer function for the pair of ϕ and θ. Next, the frame sequence of the acoustic transfer function $\hat{H}_{train}^{(\phi,\theta)}$ is estimated from $O_{train}^{(\phi,\theta)}$ using phoneme HMMs of clean speech. Then, the frame sequence of the estimated acoustic transfer function $\hat{H}_{train}^{(\phi,\theta)}$ is trained for each pair of the user's position and head orientation by SVM. For test data $O_{test}^{(\phi,\theta)}$ (any utterance), the acoustic transfer function $\hat{H}_{test}^{(\phi,\theta)}$ is estimated in the same way as the training data using a label sequence obtained from phoneme recognition. The talker position and head orientation $(\hat{\phi},\hat{\theta})$ pair is estimated by discrimination of the acoustic transfer function based on SVM.

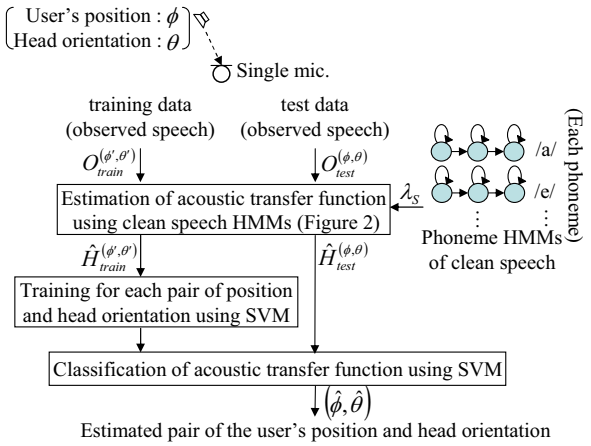

Figure 1: System overview

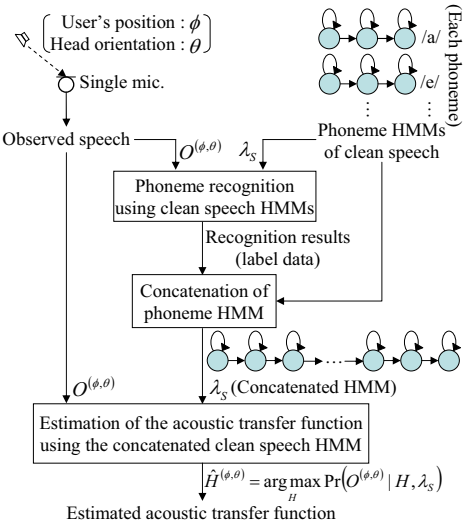

Figure 2: Estimation of the acoustic transfer function using phoneme HMMs of clean speech

Figure 2 shows the detail of the estimation of the acoustic transfer function using phoneme HMMs of clean speech. In advance, the phoneme HMMs of clean speech are trained using a clean speech database. Next, the phoneme sequence of the reverberant speech data is recognized by using each phoneme HMM of clean speech data. Using the recognition results, the phoneme HMMs are concatenated, and the frame sequence of the acoustic transfer function $\hat{H}^{(\phi,\theta)}$ is estimated from the reverberant speech $O^{(\phi,\theta)}$ based upon a maximum-likelihood (ML) estimation approach using the concatenated HMM.

2.2. Estimation of the Acoustic Transfer Function

This section presents the method for estimating the frame sequence of the acoustic transfer function [5]. The estimation is implemented by maximizing the likelihood of the observed speech data from a user's position. The reverberant speech signal in a room environment is approximately represented in the cepstral domain as

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

where O_{cep}, S_{cep}, and H_{cep} are cepstra for the reverberant speech signal, clean speech signal, and acoustic transfer function 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. 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 HMMs.

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:

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

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 written as

$$\begin{aligned} & \Pr(O,p,b_p,c_p|\hat{H},\lambda_S) \\ &= \textstyle\prod_n a_{b(n-1),b(n)} w_{b(n),c(n)} \\ &\quad \cdot N(O(n); \mu^{(S)}_{p,j,k} + \hat{H}_(n), \Sigma^{(S)}_{p,j,k}) \end{aligned} \tag{5}$$

where n, a and w represent the frame, the transition probability and the mixture weight, respectively. $N(O;\mu,\Sigma)$ denotes the multivariate Gaussian distribution, and $\mu^{(S)}_{p,j,k}$ and $\Sigma^{(S)}_{p,j,k}$ are the mean vector and the (diagonal) covariance matrix to mixture k of state j in the concatenated clean speech HMM, respectively. (4) is expanded and we focus only on the term involving H.

$$\begin{aligned} & Q(\hat{H}|H) \\ &= -\textstyle\sum_p \sum_j \sum_k \sum_n \gamma_{p,j,k}(n) \\ &\quad -\textstyle\sum_{d=1}^{D} \left\{ \frac{1}{2}\log(2\pi)^D \sigma^{(S)^2}_{p,j,k,d} \right. \\ &\quad \left. + \frac{(O(d;n)-\mu^{(S)}_{p,j,k,d}-\hat{H}(d;n))^2}{2\sigma^{(S)^2}_{p,j,k,d}} \right\} \end{aligned} \tag{6}$$

$$\gamma_{p,j,k}(n) = \Pr(O(n),p,j,k|\lambda_S) \tag{7}$$

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.

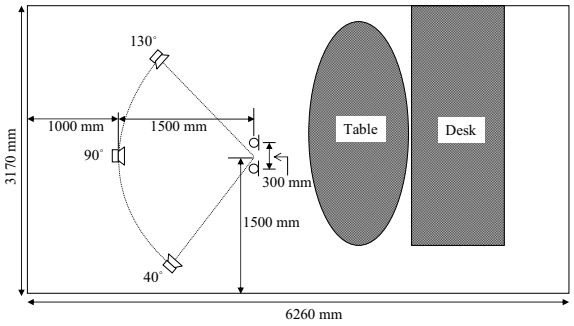

Figure 3: Experimental room environment and the loudspeaker position

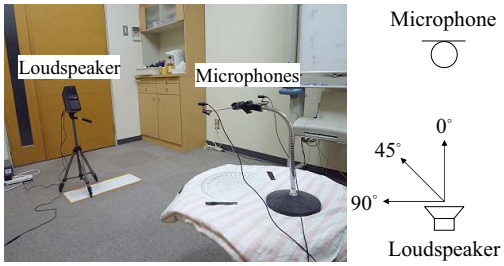

Figure 4: Photo of the recording environment and the head orientation of the loudspeaker

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

$$\hat{H}(d;n) = \frac{\sum_p \sum_j \sum_k \gamma_{p,j,k}(n) \frac{O(d;n) - \mu^{(S)}_{p,j,k,d}}{\sigma^{(S)2}_{p,j,k,d}}}{\sum_p \sum_j \sum_k \frac{\gamma_{p,j,k}(n)}{\sigma^{(S)2}_{p,j,k,d}}}. \quad (8)$$

3. Experiments

3.1. Experiment Conditions

The proposed method was evaluated in a real room environment. Figure 3 shows the experimental room environment and the position of the loudspeaker. Figure 4 depicts the recording environment and shows the head orientation of the loudspeaker. The size of the recording room was about 6.3 m × 3.2 m × 2.8 m (width × depth × height). The reverberation time was about 350 msec, and the distance to the microphone was about 2 m. The speech signal was recorded by two microphones, and the signal recorded by one of the microphones was used for the proposed method. The microphone was a directional type (SONY ECM-66B). There were three positions (40, 90 and 130 degrees) and three head orientations (0, 45 and 90 degrees) for the loudspeaker for training and testing. A total of 9 pairs (3 × 3) for position and head orientation exist. One loudspeaker (BOSE Mediamate II) was used for each position and head orientation.

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 uttered by a male 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 number of data used to train the acoustic transfer function for one pair of the position and head orientation was 50 words. The test data for one pair consisted of 166 words. The estimation accuracy was calculated by 4-fold cross-validation. 16-order MFCCs (Mel-Frequency Cepstral Coefficients) were used as feature vectors. The speech data for training the clean speech model, training the acoustic transfer function, and testing were spoken by the same person but had different text utterances, respectively. We used SVM^{light} for the Support Vector Machine with the RBF (Gaussian) kernel. Then, SVM was extended using the one-vs-rest method in order to carry out multi-class classification. For each test data (word), the position and head orientation are classified by the multi-class SVM.

Table 1: Localization accuracy of the proposed method and the median of the output direction from CSP analysis for each position.

Position	40 deg.	90 deg.	130 deg.	average
Accuracy [%]	83.6	93.5	98.3	91.8
CSP [deg.]	40.9	90	131.4	-

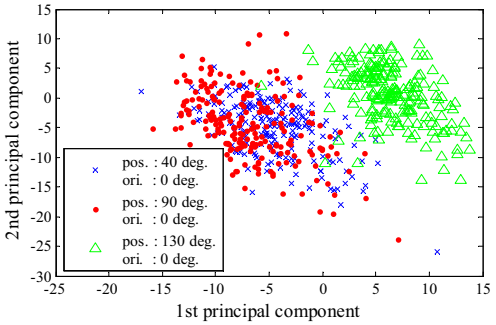

Figure 5: Mean values of the acoustic transfer function for each position fixing the head orientation at 0 deg.

3.2. Experimental Results

At first, we confirmed the performance of the proposed method in the talker localization, fixing the head orientation of the loudspeaker at 0 degrees. Table 1 shows the localization accuracy of our proposed method and the median of the output direction from CSP (Cross-power Spectrum Phase) analysis [6] for each position of the loudspeaker. CSP analysis is also known as Generalized Cross-Correlation PHAse Transform (GCC-PHAT). In CSP analysis, the time delay between the signals observed by the two microphones was estimated by searching the peak of the CSP coefficient.

As shown in this table, there is a difference in the localization accuracy between the positions of the speaker, while the CSP analysis was able to estimate the direction stably. Figure 5 shows the mean values of the acoustic transfer function for each word at three positions. The acoustic transfer functions are calculated by (2), and the total number of dimensions was reduced to two using Principal Component Analysis. As shown in this figure, the acoustic transfer function distribution for 130 degrees is easily discriminated. On the other hand, it is relatively difficult to discriminate the distribution for 40 degrees.

Next, we evaluated the performance of the proposed

Table 2: Head orientation estimation accuracies for each fixed position (pos.), where the number of head orientations (ori.) is two (0 and 90 deg.) and three (0, 45 and 90 deg.)

pos. \ ori.	0 deg.	90 deg.	average
40 deg.	81.6	84.3	83.0
90 deg.	96.1	92.6	94.4
130 deg.	94.4	93.7	94.1
average	90.7	90.2	90.5

pos. \ ori.	0 deg.	45 deg.	90 deg.	average
40 deg.	73.0	20.0	86.7	59.9
90 deg.	97.1	10.2	90.1	65.8
130 deg.	82.8	33.7	97.1	71.2
average	84.3	21.3	91.3	65.7

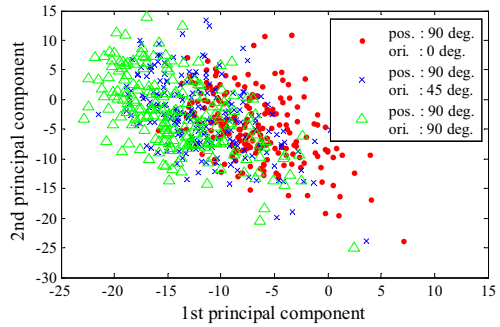

Figure 6: Mean values of the acoustic transfer function for each head orientation fixing the location at 90 deg.

method for the talker's head orientation, fixing the position of the loudspeaker for each location. Table 2 shows the head orientation estimation accuracies, where the number of head orientations is two (0 and 90 deg.) and three (0, 45 and 90 deg.). As shown in these tables, the proposed method was able to estimate the head orientation with an accuracy of over 84 %, when the head orientation was 0 degrees or 90 degrees. However, it is difficult for the proposed method to estimate a head orientation of 45 degrees. Figure 6 shows the mean values of the acoustic transfer function for each head orientation, where the speaker location is fixed at 90 degrees. As shown in this figure, the difference of the distributions for every head orientation is not as clear as that for location, and the distribution for 45 degrees, in particular, is difficult to discriminate from the other head orientations. Table 3 shows the median of the output direction from CSP analysis for each position and the head orientation. These results show that the difference in the head orientation slightly influenced the results of the CSP algorithm.

Finally, we evaluated the performance of the proposed method for both talker localization and head orientation. Table 4 shows the localization and head orientation estimation accuracy, where the number of head orientations is two and three. As shown these tables, there is also a difference in the accuracy between positions of the speaker and the accuracy for the head orientation of 45 degrees is also low. However, the proposed method was able to estimate the location and head orientation with the averaged accuracy of about 80 %, where the number of head orientations was two, and 60 %, where the number was three.

Table 3: The median of the output direction from CSP analysis for each position and the head orientation.

pos. \ ori.	0 deg.	45 deg.	90 deg.
40 deg.	40.9	40.9	40.9
90 deg.	90	90	90
130 deg.	131.4	131.4	100.9

Table 4: Localization and head orientation estimation accuracy, where the number of head orientations is two (0 and 90 deg.) and three (0, 45 and 90 deg.)

pos. \ ori.	0 deg.	90 deg.	average
40 deg.	48.6	70.8	59.7
90 deg.	87.2	93.4	90.3
130 deg.	95.0	84.5	89.8
average	77.0	82.9	79.9

pos. \ ori.	0 deg.	45 deg.	90 deg.	average
40 deg.	44.3	15.7	68.2	42.7
90 deg.	83.7	29.8	84.9	66.2
130 deg.	76.8	50.8	87.5	71.7
average	68.3	32.1	80.2	60.2

4. Conclusions

This paper has described a talker localization and head orientation estimation method using a single microphone based on discrimination of the acoustic transfer function. The sequence of the acoustic transfer function is estimated by phoneme HMMs of clean speech. The experiment results in a real room environment confirmed its effectiveness for location and head orientation estimation tasks. But the localization accuracy decreases as the number of training positions or head orientations increases. Therefore, we will research the optimal modeling of the reverberant speech and also the feature vector that retains useful information to discriminate both the acoustic transfer function for each position and head orientation. Future work will include efforts to compare our results with other published works using a network of microphone arrays.

5. Acknowledgment

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

6. References

[1] A. Brutti, M. Omologo, and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," in *Proc. Interspeech* 2005, pp. 2337-2340, 2005.

[2] J. M. Sachar and H. F. Silverman, "A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array," in *Proc. ICASSP* 2004, vol. 4, pp. 65-68, 2004.

[3] C. Segura, A. Abad, J. Hernando and C. Nadeu, "Speaker orientation estimation based on hybridation of GCC-PHAT and HLBR," in *Proc. Interspeech* 2008, pp. 1325-1328, 2008.

[4] M. Togami and Y. Kawaguchi, "Head orientation estimation of a speaker by utilizing kurtosis of a DOA histogram with restoration of distance effect," in *Proc. ICASSP* 2010, pp. 133-136, 2010.

[5] R. Takashima, T. Takiguchi and Y. Ariki, "HMM-based Separation of Acoustic Transfer Function for Single-channel Sound Source Localization," in *Proc. ICASSP* 2010, pp. 2830-2833, 2010.

[6] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based techniques," in *Proc. ICASSP* 1994, vol. 2, pp. 273-276, 1994.