Estimation of Talker's Head Orientation Based on Discrimination of the Shape of Cross-power Spectrum Phase Coefficients

Ryoichi Takashima, Tetsuya Takiguchi, 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 2-channel microphones. In recent research, some approaches based on a network of microphone arrays have been proposed in order to estimate the talker's head orientation. In those methods, the talker's head orientation is estimated using the sound amplitude or peak value of CSP (Cross-power Spectrum Phase) coefficients obtained from each microphone array. However, microphone array network systems need many microphone arrays to be set along the walls of a given room so that sub-microphone arrays surround the user. In this paper, we focus on the shape of the CSP coefficients affected by the reverberation, which depends on the talker's position and the head orientation. In our proposed method, we use not only the peak value but also the other values of the CSP coefficients as feature vectors, and the talker's position and the head orientation are estimated by discriminating the CSP vector. The effectiveness of this method has been confirmed by talker localization and head orientation estimation experiments performed in a real environment.

Index Terms: microphone array, talker localization, head orientation estimation, acoustic transfer function, CSP coefficients

1. Introduction

For human-human or human-computer interaction, the talker's location is an important cue that determines who is talking. This information can be helpful, especially in multi-user conversation scenarios such as a meeting system, robotic communication, and so on. There have been studies for understanding of a conversation scene based on the talker localization approach (e.g., [1, 2]). An approach using the turn-taking information obtained from DOA (Direction-of-Arrival) estimation results for the discrimination of system requests or users' conversations has also been proposed [3]. For more advanced understanding of the conversation scene, the talker's head orientation may also be important because it can determine not only who is talking but also who he/she is talking to. This who-talks-to-whom information is beneficial, particularly under conditions in which multiple users are having a conversation.

There have been many studies about sound source localization. On the other hand, recently, head orientation estimation from speech signals has created much interest, and some approaches have been described in [4, 5, 6, 7]. These methods use a network of microphone arrays in order to estimate the talker's head orientation. The approach described in [4] is based on the CSP algorithm, which is often used for talker localization. In that paper, they modify the CSP 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 [5]. Segura proposes techniques based on both of CSP and the radiation pattern of the magnitude and shows higher performance by combining these approaches [6]. An approach using the DOA histogram made from the DOA estimation results has also been proposed [7]. 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. Therefore, techniques based on one microphone array are of interest, especially in small-device-based scenarios.

In our previous work, we focused on the fact that the characteristics of the acoustic transfer function (i.e., reverberation) depend on the position of the sound source, and we discussed a single-channel sound source localization method based upon the discrimination of the acoustic transfer function [8]. We also proposed a single-channel head orientation estimation method based on the same framework because the acoustic transfer function may depend on not only the talker's location but also head orientation [9]. In this method, phoneme HMMs (Hidden Markov Models) of clean speech are used to separate the acoustic transfer function from observed speech at the user's position and head orientation, where the separation is performed by employing an approach based upon maximum likelihood estimation. Using the separated acoustic transfer function, the user's position and head orientation are trained with SVM (Support Vector Machine), and for each test utterance, the user's position and head orientation are estimated by discriminating the separated acoustic transfer function in the same way. However, this method cannot separate the acoustic transfer function completely, and some phonetic components still remained as noise components in the separated acoustic transfer function. For this reason, it is difficult for this method to discriminate small differences in head orientation.

In order to overcome that problem, this paper proposes a talker's head orientation estimation method using 2-channel microphones based on the discrimination of the CSP coefficients. CSP analysis has been used in many studies for source localization and head orientation estimation. However, in those studies, only the peak value of the CSP coefficients is used because they focus on the direct wave from the sound source. In our proposed method, on the other hand, we focus on the shape of the CSP coefficients affected by the reverberation in order to characterize the reverberation (acoustic transfer function), which depends on the talker's position and head orientation. Not only the peak value but also the other values of the CSP coefficients are used as feature vectors, and the talker's position and the head orientation are estimated by discriminating the CSP vector. Because CSP coefficients are normalized by the power spectrum of observed speech, the performance of this method may be more

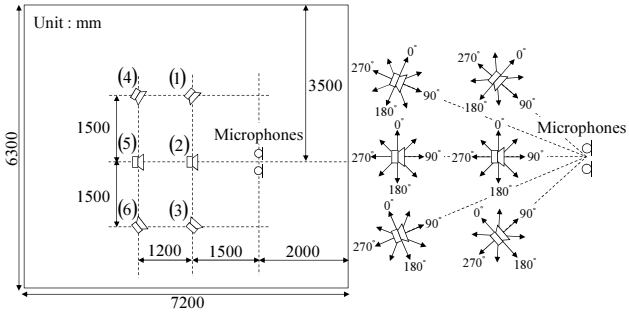

Figure 1: Experimental room environment and head orientation of a loudspeaker for each position. Each parenthetic number shows the index of the position of the loudspeaker.

robust against differences in the phoneme sequence uttered by a speaker than our previous method.

Unlike 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 the microphones 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

CSP analysis is one of the most popular methods for estimation of sound source direction and sound source localization, and it is also known as the Generalized Cross-Correlation PHAse Transform (GCC-PHAT). The CSP coefficients are obtained by applying the whitening on signals observed from two microphones and calculating their cross-correlation, as shown below.

$$CSP(\tau) = DFT^{-1}\left\{\frac{DFT(o_l(t))\cdot DFT^*(o_r(t))}{|DFT(o_l(t))|\cdot|DFT(o_r(t))|}\right\} \qquad (1)$$

$o_l(t)$ and $o_r(t)$ are discrete time sequences observed by the left channel and right channel, and τ is the time-lag of these signals, respectively. In the conventional sound source localization technique and recent head orientation estimation methods, the talker's location and head orientation are estimated by using only the peak value of the CSP coefficients.

In the case of reverberant speech, the reflected waves also cause a certain level of correlation energy in some CSP coefficients other than the peak value caused by the direct wave. For this reason, in our proposed method, all CSP coefficients are used as the feature vector in order to deal with the characteristics of the reverberation. Because CSP coefficients are normalized by the power spectrum of observed speech, the performance of this method may be more robust against differences in the phoneme sequence uttered by a speaker than our previous method.

First, we record some reverberant speech data uttered from each position with each head orientation using 2-channel microphones in order to train the position and head orientation. Next, for each training data, the CSP coefficients are calculated. Then, the CSP coefficients are trained for each pair of the user's position and head orientation using SVM. For test data (any utterance), the CSP coefficients are calculated, and the talker's position and head orientation pair is estimated by discrimination of the CSP coefficients based on SVM.

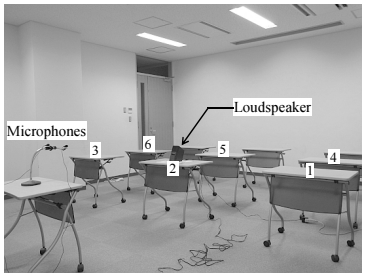

Figure 2: Photo of the recording environment. Each number shows the index of the position of the loudspeaker.

3. Experiments

3.1. Experiment Conditions

The proposed method was evaluated in a real room environment. Figure 1 shows the experimental room environment and the head orientation of loudspeaker for each position. The size of the recording room was about 7.2 m $\times$ 6.3 m $\times$ 2.8 m (width $\times$ depth $\times$ height). The reverberation time was about 1,220 msec. One loudspeaker was set at each position with each orientation, and for each position and orientation, the speech signal uttered by a male was played and recorded using two microphones. The distance between the microphones was 30 cm. The microphones were a directional type (SONY ECM-66B), and a BOSE Mediamate II was used for the loudspeaker.

There were six positions, and each parenthetic number in figure 1 shows the index of the position of the loudspeaker. In the following paragraph, each position will be defined by this index. There were eight orientations in steps of 45 degrees. The loudspeaker's orientation toward the microphones was defined as 90 degrees. A total of 48 pairs (6 $\times$ 8) for position and head orientation exist. Figure 2 depicts the recording environment. Each number shows the index of the position of the loudspeaker.

The experiment utilized the word data uttered by a male and stored in the ATR Japanese speech database. The speech signal was sampled at 12 kHz and windowed with a 32-msec Hamming window every 8 msec. Then, 512 dimensional CSP coefficients were computed for each frame, and the mean vector was used as the feature vector of one word. For each position and head orientation, 50 words were recorded and 10 of them were used for testing. The other 40 words were used for training by changing the number of training data for 1, 5, 10, 20, 30 and 40 words. The speech data for training and testing were spoken by the same speaker but used different text utterances, respectively. Changing the data set used for testing and training, the estimation accuracy was calculated by 5-fold cross-validation. The total number of test data was 2400 (50 $\times$ 48). We used SVM^{light} [10] for the Support Vector Machine with the RBF (Gaussian) kernel. Then, SVM was extended by the one-vs-rest method in order to carry out multi-class classification. For each test data (word), the position and head orientation of the speaker were classified with the multi-class SVM.

3.2. Experimental Results

Table 1 shows the localization and head orientation estimation accuracy for each number of training data. As shown in this table, the position and head orientation were estimated with an accuracy of more than 90 %, when the number of training data was more than 5 words. They were also estimated with the

Table 1: Localization and head orientation estimation accuracy for each number of training data

Number of training data (words)	1	5	10	20	30	40
Accuracy [%]	56.2	94.0	98.4	99.7	99.9	99.9

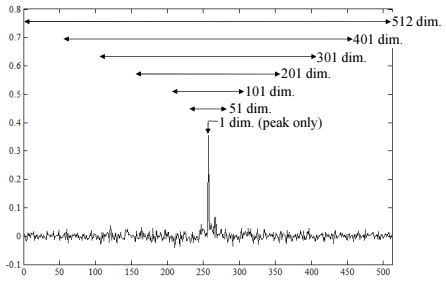

Figure 3: Number of dimensions and the range of the CSP coefficients used for the feature vector

Table 2: Head orientation accuracy [%] for each number of dimensions (dim.) and training data (num.)

dim. \ num.	1	5	10	20	30	40
1 (peak only)	22.0	24.3	24.5	26.3	26.8	22.5
51	72.3	70.8	78.8	95.8	93.0	93.0
101	82.5	94.5	97.5	87.8	97.5	95.5
201	91.3	92.8	96.5	99.5	99.5	99.0
301	92.5	94.0	95.8	99.3	99.3	99.5
401	92.5	98.3	95.3	99.3	99.5	99.5
512	91.8	97.8	95.5	99.5	99.5	99.5

accuracy of nearly 100 %, when the number of training data was more than 30 words.

In our proposed method, all CSP coefficients are used for the feature vector, while the conventional source localization techniques and recent head orientation estimation approaches focus on only the peak value of the CSP coefficients. We investigated what range of all the CSP coefficients works well for discriminating the head orientation. As shown in Figure 3, the number of dimensions of the feature vector was changed for the case of using only the peak value, peak value with the neighboring 50, 100, 200, 300 and 400 dimensions, and all CSP coefficients (512 dimensions). Table 2 shows the head orientation estimation accuracy for each number of dimensions and training data. In this experiment, only the head orientation estimation performance was evaluated by fixing the position of the loudspeaker at location 2. This was an 8-class discrimination task.

As shown in this table, the performance was improved by adding the neighboring 50 dimensions as compared with the case of using only the peak value. The performance improved as the number of dimensions increased, but there was not a drastic improvement when more than 200 dimensions were used. Also, it was difficult to discriminate the head orientation when only the peak value was used. Table 3 shows the confusion matrix for the case of using only the peak value, where the number of training data was 40. As shown in this table, test data of 90° were discriminated with an accuracy of more than 90 %. However, most of test data for 45° and 135° were faultily discriminated as 90°, which was similar to their orientations. Also the test data of the other orientations tended to be discriminated

Table 3: Accuracy confusion matrix for the case using only the peak value, where the number of training data was 40

		Predicted							
	degree	**0**	**45**	**90**	**135**	**180**	**225**	**270**	**315**
Actual	**0**	0	0	20	0	0	60	0	20
	45	0	8	74	0	0	10	0	8
	90	0	0	92	0	0	2	0	6
	135	0	8	60	0	0	22	0	10
	180	0	0	20	0	0	60	0	20
	225	0	0	20	0	0	60	0	20
	270	0	0	20	0	0	60	0	20
	315	0	0	20	0	0	60	0	20

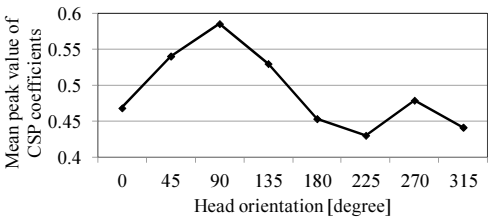

Figure 4: Mean peak value of the CSP coefficients for each head orientation

Table 4: Localization and head orientation estimation accuracy in noisy environments

SNR [dB]	5	10	20	clean
Accuracy [%]	2.2	3.1	47.5	99.9

as 225°. Figure 4 shows the mean peak value of the CSP coefficients for each head orientation. As shown in this figure, the CSP coefficients of 45°, 90° and 135° had high peak values, and those of the other orientations had low peak values.

These results indicate that when only the peak value was used, the head orientations could be discriminated based only upon the criterion that the peak value of its CSP coefficients was high or low. The recent approaches utilize some microphone arrays and estimate the talker's head orientation by comparing the peak values of CSP coefficients calculated by each microphone array. However, when only a 2-channel microphone array is utilized, the head orientation cannot be discriminated using only the peak value of the CSP coefficients, and adding the other dimension of CSP coefficients enables it to be discriminated.

Next, we added a pink noise to speech data recorded using the 2-channel microphones, and evaluated the robustness against the noise environment. Among recorded speech data, only the test data had the same pink noise added so that SNR (Signal to Noise Ratio) was 5 dB, 10 dB and 20 dB. Table 4 shows the localization and head orientation estimation accuracy for each SNR. The number of training data was 40. As shown in this table, the performance decreased drastically by adding the pink noise. In the case of the SNRs for 5 dB and 10 dB, the localization and head orientation accuracy were 2.2 % and 3.1 %, respectively, which is, more or less, the same as the expected value obtained by discriminating the 48 classes randomly (2.1 %).

Table 5 shows the localization and head orientation estimation accuracy calculated for each position of the loudspeaker, at the SNR of 20 dB. As shown in this table, the performance

Table 5: Localization and head orientation estimation accuracy for each loudspeaker position at 20 dB SNR

Location	1	2	3	4	5	6	mean
Accuracy [%]	61.3	18.3	59.3	60.5	24.8	61.0	47.5

Table 6: Accuracy confusion matrix for SNR of 20 dB, where the loudspeaker position was fixed at location 2

		Predicted							
	degree	**0**	**45**	**90**	**135**	**180**	**225**	**270**	**315**
Actual	**0**	4	0	96	0	0	0	0	0
	45	0	2	98	0	0	0	0	0
	90	0	0	100	0	0	0	0	0
	135	0	0	100	0	0	0	0	0
	180	0	0	98	0	2	0	0	0
	225	0	0	96	0	0	4	0	0
	270	0	0	98	0	0	0	2	0
	315	0	0	96	0	0	0	0	4

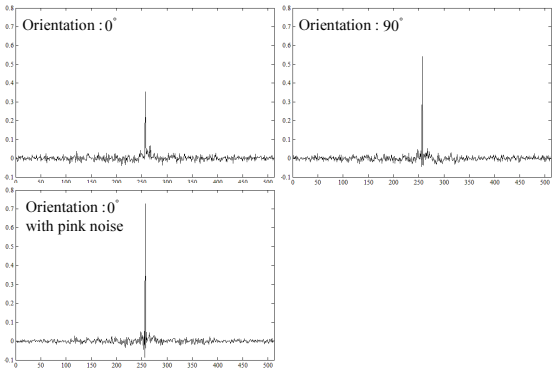

Figure 5: CSP coefficients at location 2. Upper left: head orientation 0°. Upper right: head orientation 90°. Lower left: head orientation 0° with pink noise was added.

at locations 2 and 5, which were in front of the microphones, were especially low. Also, we evaluated only the head orientation estimation performance for the SNR of 20 dB fixing the position of the loudspeaker at location 2. The head orientation estimation accuracy was 14.8 %. Table 6 shows the confusion matrix for this evaluation. As shown in this table, almost all the test data were discriminated as 90°.

This is because the same pink noise was added to the speech data recorded by both channels, and the CSP coefficients had a high coefficient energy at the point where the phase delay was 0. Figure 5 shows the CSP coefficients at location 2. The figure at the upper left shows the CSP coefficients for the head orientation of 0°, and the figure at the upper right shows those of 90°, where noise was not added to the test data. The figure at the lower left shows those of noise-added data for the head orientation of 0°.

As shown in these figures, the peak value at the middle point of the CSP coefficients for 0° increased due to the correlation between the same noise added to the left and right channels. As a result, these noise-added CSP coefficients were discriminated as 90°. For the case of the SNRs of 5 dB and 10 dB, the value at the middle point increased further. Therefore, even the speech uttered from a location other than 2 and 5, which did not have the peak at the middle point essentially, were discriminated as location 2 and with a head orientation of 90°, and this is why the performance decreased drastically.

4. Conclusion

This paper has described a talker localization and head orientation estimation method using 2-channel microphones based on discrimination of the CSP coefficients. The characteristics of the reverberation, which depends on the talker's location and head orientation, are represented by CSP coefficients. The talker's position and head orientation are estimated by discriminating the CSP coefficients with SVM. The experiment results in a real room environment show that our proposed method can discriminate 6 positions and 8 head orientations with a maximum accuracy of about 99 %.

Also, the results of the comparison experiment, which used only the peak value of the CSP coefficients, show that the use of the dimensions near the peak value enables us to estimate the head orientation using only 2-channel microphones, while recent approaches need a number of microphone arrays. However, the performance decreases drastically when noise is added. In order to overcome this problem, this system will need to be able to discriminate the sound source as either speech or noise. Future work will include efforts to investigate the performance when the recording environment changes. We will also research the estimation for unknown positions or head orientations.

5. Acknowledgment

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

6. References

[1] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, pp. 2011–2022, 2007.

[2] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "A DOA based speaker diarization system for real meetings," in *HSCMA2008*, 2008, pp. 29–32.

[3] T. Yamagata, A. Sako, T. Takiguchi, and Y. Ariki, "System request detection in conversation based on acoustic and speaker alternation features," in *Proc. Interspeech07*, 2007, pp. 2789–2792.

[4] 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. Interspeech05*, 2005, pp. 2337–2340.

[5] 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. ICASSP04*, 2004, vol. 4, pp. 65–68.

[6] C. Segura, A. Abad, J. Hernando, and C. Nadeu, "Speaker orientation estimation based on hybridation of GCC-PHAT and HLBR," in *Proc. Interspeech08*, 2008, pp. 1325–1328.

[7] 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. ICASSP10*, 2010, pp. 133–136.

[8] R. Takashima, T. Takiguchi, and Y. Ariki, "HMM-based separation of acoustic transfer function for single-channel sound source localization," in *Proc. ICASSP10*, 2010, pp. 2830–2833.

[9] R. Takashima, T. Takiguchi, and Y. Ariki, "Single-channel head orientation estimation based on discrimination of acoustic transfer function," in *Proc. Interspeech11*, 2011, pp. 2721–2724.

[10] T. Joachims, *Making large-scale SVM learning practical*, MIT Press, 1999.