FEATURE SELECTION BASED ON MULTIPLE KERNEL LEARNING FOR SINGLE-CHANNEL SOUND SOURCE LOCALIZATION USING THE ACOUSTIC TRANSFER FUNCTION

Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki

Graduate School of System Informatics, 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. In our previous work [1], we discussed the single-channel sound source localization method, where the acoustic transfer function from a user's position is estimated by using a Hidden Markov Model (HMM) of clean speech in the cepstral domain. In this paper, each cepstral dimension of the acoustic transfer function is newly selected in order to select the cepstral dimensions having information that is useful for classifying the user's position. Then, we propose a feature selection method for the cepstral parameter using Multiple Kernel Learning (MKL) to define the base kernels for each cepstral dimension (scalar) of the acoustic transfer function. The user's position is trained and classified by Support Vector Machine (SVM). The effectiveness of this method has been confirmed by sound source (talker) localization experiments performed in a room environment.

Index Terms— single channel, talker localization, feature selection, maximum likelihood, Multiple Kernel Learning

1. INTRODUCTION

Many systems using microphone arrays have been tried to localize sound sources. Conventional techniques, such as MUSIC, CSP, and so on (e.g., [2, 3]), 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., [4, 5]). 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., [6, 7]). In our previous work [1], 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 a 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 Gaussian Mixture Model (GMM) of the acoustic transfer function is trained to deal with the influence of a room impulse response. Then, for each test utterance, we find a maximum-likelihood GMM from among the estimated GMMs corresponding to each position.

In each cepstral dimension of the acoustic transfer function, however, some dimensions may be strongly affected by the impulse response of the user's position, and others may be affected only minimally. In this paper, each cepstral dimension of the acoustic transfer function is newly selected in order to select the cepstral dimensions having useful information for classifying user's position. Then, we propose a feature selection method for the cepstral parameter using Multiple Kernel Learning (MKL) [8] defining the base kernels for each cepstral dimension (scalar) of the acoustic transfer function. The user's position is trained and classified by SVM. 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 system overview. First, we record the reverberant speech data $O_{train}^{(\theta)}$ from each position θ in order to train the acoustic transfer function for θ. Next, the frame sequence of the acoustic transfer function $\hat{H}_{train}^{(\theta)}$ is estimated from $O_{train}^{(\theta)}$ using phoneme HMMs of clean speech. Then, the cepstral parameters of estimated acoustic transfer function $\hat{H}_{train}^{(\theta)}$ and the feature weights are trained for each user's position θ by MKL-SVM. For test data $O_{test}^{(\theta)}$ (any utterance), the acoustic transfer function $\hat{H}_{test}^{(\theta)}$ is estimated in the same way as the training data using a label sequence obtained from a phoneme recognition. The talker position $\hat{\theta}$ is estimated by discrimination of the acoustic transfer function based on SVM.

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 $\hat{H}^{(\theta)}$ is estimated from the reverberant speech $O^{(\theta)}$ based upon a maximum-likelihood (ML) estimation approach using the concatenated HMM.

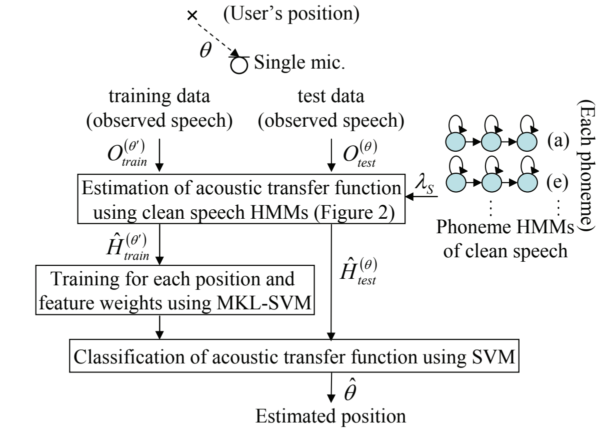
(User's position)
θ
Single mic.
training data
(observed speech)
test data
(observed speech)
$O^{(\theta')}_{train}$
$O^{(\theta)}_{test}$
(a)
(e)
(Each phoneme)
Estimation of acoustic transfer function
using clean speech HMMs (Figure 2)
λ_S
Phoneme HMMs
of clean speech
$\hat{H}^{(\theta')}_{train}$
$\hat{H}^{(\theta)}_{test}$
Training for each position and
feature weights using MKL-SVM
Classification of acoustic transfer function using SVM
$\hat{\theta}$
Estimated position

Fig. 1. System overview

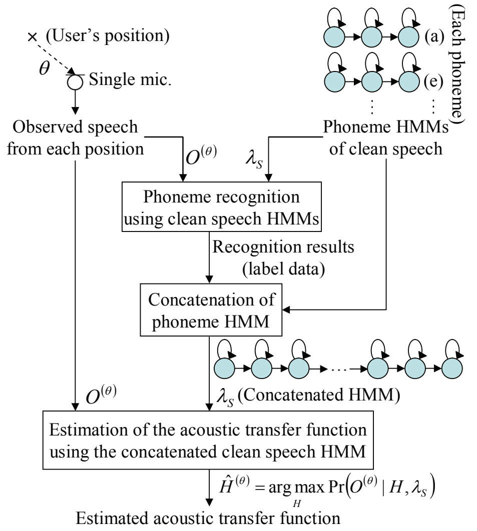

Fig. 2. Estimation of the acoustic transfer function using phoneme HMMs of clean speech

2.2. Maximum-Likelihood-Based Parameter Estimation

This section presents the method for estimating the frame sequence of the acoustic transfer function [1]. 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:

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

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.

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

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

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

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.

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

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

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.

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

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

3. FEATURE SELECTION AND CLASSIFICATION USING MKL-SVM

In our previous work, using the estimated frame sequence data of the acoustic transfer function, the GMM for the acoustic transfer function was trained for each user's position. For test data, the talker position was estimated by finding a GMM having the maximum-likelihood from among the estimated GMMs corresponding to each position. In each cepstral dimension of the acoustic transfer function, however, some dimensions may be strongly affected by the impulse response of the user's position, and others may be affected only minimally. In this paper, each cepstral dimension of the acoustic transfer function is newly selected by using Multiple Kernel Learning (MKL) [8] in order to select the cepstral dimensions having information that is useful for classifying the user's position. Then, the estimated acoustic transfer function for each position is classified by SVM.

In a MKL framework, a combined kernel function is defined as a linear combination of the base kernel.

$$k(\mathbf{H}_i, \mathbf{H}_j) = \textstyle\sum_l \beta_l k_l(\mathbf{H}_i, \mathbf{H}_j) \tag{9}$$

Here k_l is the l-th base kernel computed from the i-th and j-th samples of the acoustic transfer function H_i and H_j, and the non-negative coefficient β_l represents the weight of the base kernel. The MKL approach for SVM is essentially used to combine classifiers of various kernels in order to improve the classifier performance. In recent image recognition research, the MKL approach is also being used for feature vector selection. In that approach, the weights of feature vectors are trained by defining the base kernels for each different feature vector [9]. In this paper, we propose a feature selection method for the cepstral parameter, where the weights of each cepstral dimension are trained by MKL, defining the base kernels for each cepstral dimension (scalar) of the acoustic transfer function.

$$k(\mathbf{H}_i, \mathbf{H}_j) = \textstyle\sum_d \beta_d k_d(H_i(d), H_j(d)) \tag{10}$$

By defining the kernels for each element of a feature vector, the information related to the correlations between the elements in the feature vector are lost. However, the cepstral parameter is a dimensionally-uncorrelated feature compressed by a discrete cosine transform. Therefore, the lost information associated with the correlations should not influence the classification performance critically. We also expect that this feature selection method may be effective for not only our talker localization task, but also various SVM-based classification tasks using dimensionally-uncorrelated features, such as the cepstral parameter.

The kernel weight β_d is trained in the SVM framework (i.e., maximum-margin based scheme). In the SVM framework, the MKL criterion is defined by the following objective function [8].

$$\begin{aligned} \max_{\alpha,\beta} \quad & \textstyle\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \sum_d \beta_d k_d(H_i(d), H_j(d)) \\ s.t. \quad & \begin{cases} \sum_i y_i \alpha_i = 0, & 0 \le \alpha_i \le C \\ \sum_d \beta_d = 1, & \beta_d \ge 0 \end{cases} \end{aligned} \tag{11}$$

Here α_i is the Lagrange coefficient, and $y_i = \{+1, -1\}$ denotes the class label of example H_i. C determines the trade-off between the margin and training data error. In (11), both α_i and β_d are optimized by a two-step iterative procedure. In the first step, β_d is fixed, and α_i is updated by a standard SVM solver. In the second step, α_i is fixed, and β_d is updated. In this paper, we use SVM^{light} to obtain α_i, and optimize β_d by a projected-gradient scheme. In this way, the feature weights and the classification boundary are trained simultaneously.

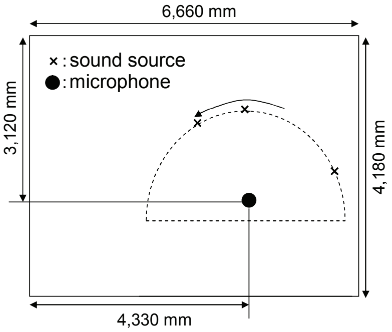

Fig. 3. Experimental room environment

4. EXPERIMENTS

4.1. Experiment Conditions

The new talker localization method was evaluated in 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 [10]. 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.

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 and the feature weights for one location was 50 words. 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 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). Then, SVM was extended by one-vs-rest method in order to carry out multi-class classification. For each test data (word), the talker position is classified by the multi-class SVM.

4.2. Experiment Results

The proposed method for classifying the acoustic transfer function using MKL-SVM was compared with our previous work [1] using 8-mix GMM and standard SVM with a single kernel. For the SVM-based method, a Gaussian kernel was employed as the kernel function, and the hyper parameter C was 1. The cepstral parameter was a dimensionally-uncorrelated feature compressed by a discrete cosine transform. As a result, there was the possibility that the optimal kernel parameters for each cepstral dimension might not be the same. For this reason, we handled MKL-SVM in two ways. One defines an identical kernel for each cepstral dimension. The other defines kernels having different kernel parameters (i.e., standard deviation of the Gaussian kernel) for each cepstral dimension. The kernel parameters were set empirically.

As shown in Figure 4, the SVM-based method showed better performance than the use of GMM. Also, our proposed methods using MKL-SVM improved the performance of the standard SVM

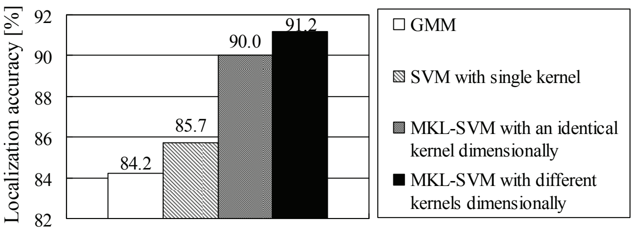

Fig. 4. Performance comparison of each classifier

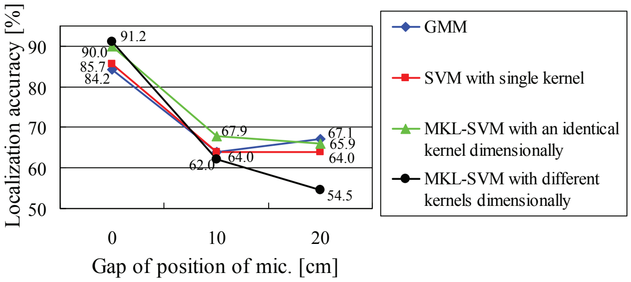

Fig. 5. Influence by the gap between microphone positions of training and test

Table 1. Feature weights for some cepstral dimensions trained using MKL for each position. **Bold type** shows the highest weight for the position.

order deg.	1^{st}	4^{th}	7^{th}	8^{th}	10^{th}
30 degrees	0.00	0.07	0.07	0.07	**0.08**
90 degrees	0.00	0.06	**0.10**	0.07	0.07
130 degrees	0.01	0.07	0.06	**0.11**	0.07

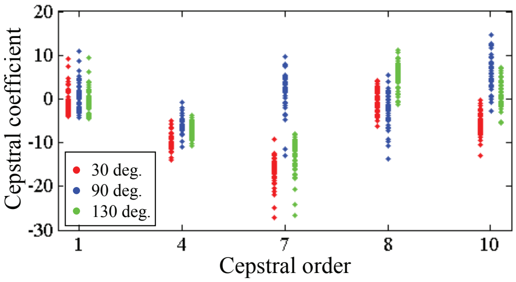

Fig. 6. Mean acoustic transfer function values for some cepstral dimensions

classifier. In addition, by defining a different parameter for each base kernel, the performance could be improved. This may show that the optimal kernel parameters for each cepstral dimension are different.

We also evaluated the performance by shifting the position of the microphone when testing and training. As shown in Figure 5, performance degraded across the board when the position of the microphone changed by 10 cm. In particular, the performance of MKL-SVM defining the different parameters for each base kernel degraded worse than other methods. This might have been because the model became sensitive to the change in the environment, while the accuracy of the model was increased by defining the kernel parameter for each cepstral dimension.

In our proposed method, the three classification boundaries, where the number of classes (positions) is three, are trained by using a one-vs-rest method. And the feature weights are trained for every classification boundary. This means that the set of feature weights is trained for each position. Table 1 shows the feature weights for some cepstral dimensions trained using MKL for each position. As shown in this table, the 10th, 7th and 8th cepstral orders have the highest weights for 30 deg., 90 deg. and 130 deg., respectively. And for every position, the 1st order has the lowest weight. Figure 6 shows the mean acoustic transfer function values for some cepstral dimensions, where the acoustic transfer functions are calculated by (2). As shown in this figure, the acoustic transfer functions for the cepstral dimensions having the highest weights distribute to be able to discriminate easily. For example, the 7th order of the acoustic transfer function at 90 degrees distributes in such a way as to be easily discriminated from those at the other positions. On the other hand, the acoustic transfer functions for the 1st order are only slightly influenced by a change in talker position.

5. CONCLUSION

This paper has described a voice (talker) localization method using a single microphone. The sequence of the acoustic transfer function is estimated by HMMs of clean speech. Then, each cepstral dimension of the acoustic transfer function is newly selected by our proposed feature selection method using MKL, defining the base kernels for each cepstral dimension. In the room environment experiment, the proposed method using MKL-SVM improved the performances of our previous work using GMM and that of standard SVM. But the localization accuracy decreases as the recording environment changes. Therefore, in the future, we will research an adaptation method in order to adapt the system to a change in room environment. In addition, we carry out research with the aim of achieving higher accuracy in the estimation of the acoustic transfer function.

6. REFERENCES

[1] R. Takashima, T. Takiguchi, and Y. Ariki, "HMM-based separation of acoustic transfer function for single-channel sound source localization," in *Proc. ICASSP2010*, 2010, pp. 2830–2833.

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

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

[4] 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.

[5] 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.

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

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

[8] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "More efficiency in multiple kernel learning," in *Proc. ICMKL*, 2007, pp. 775–782.

[9] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. ICCV07*, 2007, pp. 1–8.

[10] 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.