PREDICTION OF UNLEARNED POSITION BASED ON LOCAL REGRESSION FOR SINGLE-CHANNEL TALKER LOCALIZATION USING 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, we discussed the single-channel sound-source localization method based on the discrimination of the acoustic transfer function. However, that method requires the training of the acoustic transfer function for each possible position in advance, and it is difficult to estimate the position that has not been pre-trained. In order to estimate such unlearned positions, in this paper, we discuss a single-channel talker localization method based on a regression model, which predicts the position from the acoustic transfer function. For training the regression model, we use the local regression approach, which trains the regression model from only training samples that are similar to the evaluation data. Considering both the linear and non-linear regression models, the effectiveness of this method has been confirmed by sound-source localization experiments performed in different room environments.

Index Terms— talker localization, acoustic transfer function, local regression, Gaussian process regression, support vector regression

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., [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]). Studies focusing on the techniques for monaural sound-source localization are also being carried out [7, 8]. In these studies, the information obtained from the external ear, such as head-related transfer functions (HRTFs), is used to localize the sound source.

In our previous work [9], we discussed a single-channel sound-source localization method based on the discrimination of the acoustic transfer function. 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 Hidden Markov Model (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 uttered from a given position, where measurement of impulse responses is not required. Using the separated acoustic transfer function, the talker's position is trained in advance. Then, for each utterance, the talker's position is estimated by discriminating the acoustic transfer function separated from the observed signal because the characteristics of the acoustic transfer function depend on each position.

That method can localize the talker using only a single microphone, without the external ear used in other monaural sound-source localization studies. However, it was difficult for that method to estimate a position that has not been pre-trained because that method was based on a pattern discrimination approach. In order to estimate such unlearned positions, in this paper, we discuss a single-channel talker localization method based on a regression model. The regression model is trained using the acoustic transfer function of limited training positions. Then, for test data, the position is predicted using the separated acoustic transfer function and the regression model even if the position has not been pre-trained.

We use Multiple Linear Regression (MLR) as the linear regression model and Gaussian Process Regression (GPR) [10] and Support Vector Regression (SVR) [11] as the non-linear regression models. In addition, for training the regression model, we use the local regression approach, which trains the regression model from only training samples that are similar to the evaluation data. The effectiveness of this method has been confirmed by sound-source localization experiments performed in different room environments.

2. PROPOSED METHOD

2.1. System Overview

Figure 1 shows the system overview. First, we record the reverberant speech data O^{train} from each training position θ^{train} in order to train the regression model. Next, the acoustic transfer function $\hat{H}^{train}$ is estimated from the reverberant training speech data O^{train} using phoneme HMMs of clean speech. Using the pair of the estimated acoustic transfer function $\hat{H}^{train}$ and the position label θ^{train}, the regression model $f(H)$ which predicts the position from the acoustic transfer function is trained. For test data O^{test} (any utterance), the acoustic transfer function $\hat{H}^{test}$ is estimated in the same way as the training data using a label sequence obtained from a phoneme recognition [9]. The talker position $\hat{\theta}$ is estimated from the acoustic transfer function using the regression model.

2.2. Estimation of the Acoustic Transfer Function

Figure 2 shows the detail of the estimation of the acoustic transfer function using phoneme HMMs of clean speech [9]. In advance, the phoneme HMMs of clean speech are trained using a clean speech database. Next, the phoneme sequence of the reverberant speech

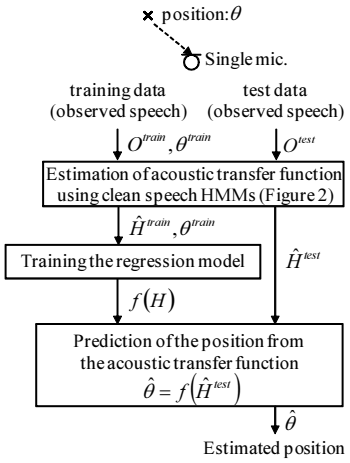

Fig. 1. System overview

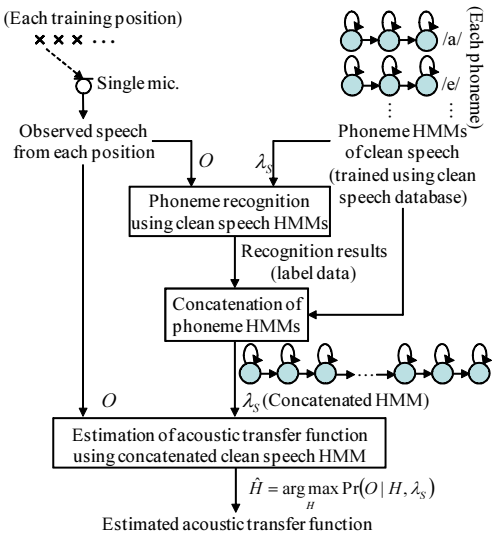

Fig. 2. Estimation of the acoustic transfer function using phoneme HMMs of clean speech

data is recognized by using each phoneme HMM of clean speech data. Using the recognition results, the phoneme HMMs are concatenated, and the acoustic transfer function $\hat{H}$ is estimated from the reverberant speech O based upon a maximum-likelihood (ML) estimation approach using the concatenated HMM.

In this method, 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 of frame n, respectively. d is the dimension of the cepstrum. 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 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 maximization step (M-step) in the EM algorithm becomes "max $Q(\hat{H}|H)$". The re-estimation formula can be derived, knowing that $\partial Q(\hat{H}|H)/\partial \hat{H} = 0$ as

$$\hat{H}(d;n) = \frac{\sum_p \sum_j \sum_k \gamma_{p,j,k}(n) \frac{O(d;n) - \mu^{(S)}_{p,j,k,d}}{\sigma^{(S)^2}_{p,j,k,d}}}{\sum_p \sum_j \sum_k \frac{\gamma_{p,j,k}(n)}{\sigma^{(S)^2}_{p,j,k,d}}}, \tag{5}$$

$$\gamma_{p,j,k}(n) = \Pr(p, j, k|O(n), H, \lambda_S). \tag{6}$$

Here $\mu^{(S)}_{p,j,k,d}$ and $\sigma^{(S)^2}_{p,j,k,d}$ are the d-th mean value and the d-th diagonal variance value of the state $b_p(n) = j$ and the mixture$c_p(n) = k$, respectively (for more details, see [9]).

2.3. Prediction of the position using the regression model

Using the estimated acoustic transfer function $\hat{H}^{train}$ of the training position and the position label θ^{train}, the regression model $f(H)$, which predicts the position from the acoustic transfer function, is trained. In this study, we use MLR as the linear regression model and GPR [10] and SVR [11] as the non-linear regression models. In addition, for training the regression model, we use the local regression approach, which trains the regression model from only training samples that are similar to the evaluation data.

2.3.1. Gaussian Process Regression

When there is a training data set, which consists of pairs of the acoustic transfer function (explanatory variable) and the position label (objective variable) $Z^{train}_n = (\hat{H}^{train}_n, \theta^{train}_n), (n = 1, \ldots, N)$, we predict the position θ^{test} of the test utterance from the acoustic transfer function $\hat{H}^{test}$. In a GPR framework, when the training set and the explanatory variable of the test data are observed, the posteriori

probability of the objective variable is assumed to be the following normal distribution:

$$\begin{aligned}\Pr(\theta^{test}|\hat{H}^{test}, Z_1^{train}, \ldots, Z_N^{train}) \\ \sim \quad \mathcal{N}(\mathbf{K}_*\mathbf{K}^{-1}\boldsymbol{\Theta}, \mathbf{K}_{**} - \mathbf{K}_*\mathbf{K}^{-1}\mathbf{K}_*^{\mathbf{T}}).\end{aligned} \tag{7}$$

Here, $\Theta = [\theta_1^{train}, \ldots, \theta_N^{train}]^T$ and $\mathbf{K}$ is a gram matrix:

$$\begin{aligned}\mathbf{K} &= \begin{bmatrix} k(\hat{H}_1^{train}, \hat{H}_1^{train}) & \cdots & k(\hat{H}_1^{train}, \hat{H}_N^{train}) \\ \vdots & \ddots & \vdots \\ k(\hat{H}_N^{train}, \hat{H}_1^{train}) & \cdots & k(\hat{H}_N^{train}, \hat{H}_N^{train}) \end{bmatrix} \\ \mathbf{K}_* &= \left[k(\hat{H}^{test}, \hat{H}_1^{train}), \cdots, k(\hat{H}^{test}, \hat{H}_N^{train})\right] \\ \mathbf{K}_{**} &= k(\hat{H}^{test}, \hat{H}^{test}).\end{aligned} \tag{8}$$

$k(H, H)$ is a kernel function and the RBF kernel [10] is used in this study. The estimated objective variable of the test data $\hat{\theta}^{test}$ is given by maximizing Eq. (7); i.e., the mean value of the normal distribution:

$$\hat{\theta}^{test} = f(\hat{H}^{test}) = \mathbf{K}_*\mathbf{K}^{-1}\boldsymbol{\Theta}. \tag{9}$$

2.3.2. Support Vector Regression

SVR is a non-linear regression using a kernel function. In a SVR framework, the objective variable is estimated using the following regression model:

$$\hat{\theta}^{test} = f(\hat{H}^{test}) = w^T\phi(\hat{H}^{test}) + b \tag{10}$$

where ϕ is a kernel mapping function (RBF kernel was used in this paper). The model parameter w and b are calculated to satisfy the following objective function.

$$\min \qquad \tfrac{1}{2}||w||^2 + C\sum_{n=1}^{N}(\xi_n + \xi_n^*) \tag{11}$$

$$s.t. \qquad \begin{cases} \theta_n^{train} - w^T\phi(\hat{H}_n^{train}) - b \le \epsilon + \xi_n \\ w^T\phi(\hat{H}_n^{train}) + b - \theta_n^{train} \le \epsilon + \xi_n^* \\ \xi_n, \xi_n^* \ge 0 \end{cases} \tag{12}$$

The first term of Eq. (11) is a regularization term, and the second one is a penalization term for the regression error ξ over the acceptable error range ϵ. C determines the trade-off between the first and second terms. In this paper, we use the SVM-KM Matlab Toolbox [12] in order to obtain the model parameter.

2.3.3. Local Regression

The regression approach assumes a correlativity between the position and the acoustic transfer function. Depending on the room environment, however, it is possible that the acoustic transfer functions are completely different even if the positions are close to each other. Considering such cases, it might be difficult to represent the acoustic transfer functions of all positions in a room using only one regression model.

Therefore, instead of a global regression model trained from all training samples, we use the local regression approach, which trains the regression model from only training samples that are similar to the evaluation data. Chao et al. [13] proposes a facial age estimation method using local regression with the SVR, and shows higher performance than the standard SVR.

Local regression is a method that combines the regression analysis and the K-nearest neighbor (K-NN) method. In this method, all training samples are held in the database. For a test sample, the K-nearest samples are picked up from the training data set. Then, the regression model for the test data is trained using only the K samples and then predicts the position.

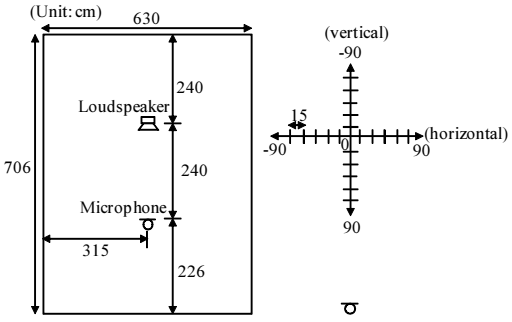

Fig. 3. Experiment room environment for sound-source localization (on horizontal and vertical axes) and the position of the loudspeaker

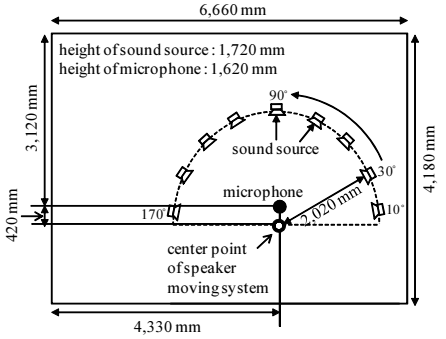

Fig. 4. Experiment room environment for the estimation of the sound-source direction (angle)

3. EXPERIMENTS

3.1. Experiment Conditions

The proposed method was evaluated in simulated reverberant environments. The reverberant speech was simulated by a linear convolution of clean speech and impulse response. The impulse responses were recorded in two different room environments.

In the experiments performed in one of the room environments, the position of a loudspeaker on a horizontal or vertical axis was estimated. Figure 3 shows the experimental room environment and the position of the loudspeaker. The reverberation time was 1,220 msec. For both the horizontal and vertical axes, training positions consisted of 7 positions (-90, -60, ..., 60, 90cm), and test positions consisted of 13 positions (-90, -75, 60, ..., 60, 75, 90cm) including 6 unlearned positions (-75, -45, -15, 15, 45, 75cm). The system estimated only the position on one axis, and the position through the other axis was fixed to 0cm (given position).

In the experiments performed in the other room environment, the position of a loudspeaker on a circular arc was estimated, where the distance to the microphone was given and the system estimated only the direction (angle) of the loudspeaker. The impulse response was taken from the RWCP database in real acoustical environments

Table 1. RMSE of the estimated position for each number of the nearest samples K used for training the local regression model ('global' shows the use of the global regression model). The number to the left of the slash shows the RMSE for unlearned positions and the right side shows that for pre-trained positions.

Acoustic transfer function H_{sub} computed using true clean speech signal											
	horizontal axis [cm]				vertical axis [cm]				angle [degree]		
	K = 50	K = 150	K = 250	global	K = 50	K = 150	K = 250	global	K = 50	K = 150	global
MLR	52.3/47.4	35.4/35.1	33.5/33.8	34.1/35.3	19.6/18.6	20.3/19.9	22.1/22.1	23.5/25.1	63.9/12.1	63.6/14.5	65.3/17.0
GPR	31.4/29.0	35.0/31.6	34.8/31.4	34.8/31.4	12.2/11.6	14.3/14.1	16.1/15.4	17.4/16.3	28.5/**10.7**	22.9/11.4	22.0/11.4
SVR	**25.6**/**26.7**	27.8/29.3	28.6/30.9	28.8/31.3	**7.7**/ **9.1**	13.3/17.8	16.8/20.8	19.4/23.7	20.3/**10.8**	**18.9**/12.1	22.7/12.2

Acoustic transfer function H_{est} estimated using clean speech HMMs											
	horizontal axis[cm]				vertical axis [cm]				angle [degree]		
	K = 50	K = 150	K = 250	global	K = 50	K = 150	K = 250	global	K = 50	K = 150	global
MLR	54.2/55.5	42.1/41.5	41.0/41.2	41.0/42.5	34.7/33.6	30.5/31.6	30.2/32.8	30.3/34.6	50.3/49.2	42.6/41.7	42.2/42.8
GPR	39.6/40.2	41.8/41.3	41.5/41.4	41.7/41.4	21.2/22.8	22.7/25.0	23.7/25.9	24.3/26.3	32.7/40.9	33.7/42.3	33.4/43.4
SVR	**36.1**/**38.0**	36.4/38.2	37.3/38.6	38.3/39.1	**17.7**/**20.3**	22.0/26.6	24.0/28.7	26.9/31.6	**29.4**/**31.5**	31.6/35.3	32.5/37.1

[14]. Figure 4 shows the experimental room environment. The reverberation time was 300 msec. The training positions consisted of 5 positions (10, 50, 90, 130, 170 degrees), and test positions consisted of 9 positions (10, 30, 50, ...,130, 150, 170 degrees) including 4 unlearned positions (30, 70, 110, 150 degrees).

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 regression model was 50 words ($\times$ number of training positions). The test data for one location consisted of 166 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 regression model, and testing were spoken by the same speakers but had different text utterances, respectively.

3.2. Experiment Results

We evaluated the performance of the method using the acoustic transfer function H_{sub} computed with Eq. (2) using true clean speech signal $S_{cep}(d;n)$ and H_{est} estimated with Eq. (5) using clean speech HMMs. Table 1 shows the Root Mean Square Error (RMSE) of the position estimated using the local regression model with K-nearest samples and the global regression model trained using all training samples.

When H_{sub} was used, the proposed method could estimate the pre-trained and unlearned positions of the loudspeaker on the vertical axis with a minimum RMSE of 9.1 and 7.7cm, respectively. In the case of the vertical axis, only the distance between the microphone and the sound source was changed, and the direction of sound source was fixed to 0 degrees. Hence, this result means that when the direction of the sound source is fixed, the distance between the microphone and the sound source can be predicted from the acoustic transfer function relatively easily. In the case of the positions on the circular arc in Fig. 4, MLR showed a much higher error on estimating the unlearned position than on estimating the pre-trained position. This result means that the direction of the sound source is difficult to represent using the linear regression model. However, the use of a non-linear regression model could decrease the prediction error.

When H_{est} was used, the performances decreased for all experimental conditions compared with that using true clean speech signals in Eq. (2). This is because the acoustic transfer function was not separated completely from the observed speech and it was influenced to some extent by the difference between the utterance texts for training and for testing. In both cases using H_{sub} and H_{est}, the local regression approach outperformed the global regression approach by using the non-linear regression method.

4. CONCLUSION

This paper has described a talker localization method using a single microphone. The acoustic transfer function is estimated using HMMs of clean speech. Then, using the acoustic transfer function, the regression model, which predicts the talker's position from the acoustic transfer function, is trained. For training the regression model, we use the local regression approach, which trains the regression model from only training samples that are similar to the evaluation data. Considering both the linear and non-linear regression models, the effectiveness of this method has been confirmed by sound-source localization experiments performed in different room environments.

The proposed method showed higher performances on the estimation of the position on the vertical axis (i.e., distance to the microphone). It is difficult for two-channel microphones to estimate the distance to the microphone even though the direction of the sound source can be estimated easily. Therefore, the proposed approach using the acoustic transfer function might improve the performance of conventional multi-microphone systems.

However, the localization errors on the horizontal axis and circular arc were higher than those on the vertical axis. In order to reduce such errors, more information about the other training position might be required. In addition, more accurate estimation of the acoustic transfer function is also important. Future work will include efforts to study the performance of the estimation of a 2-D position, considering both the horizontal and the vertical axes.

5. ACKNOWLEDGMENT

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

6. REFERENCES

[1] D. Johnson and D. Dudgeon, *Array Signal Processing*, Prentice Hall, Upper Saddle River, NJ, USA, 1996.

[2] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using CSP analysis," in *Proc. ICASSP96*, Atlanta, Ga, USA, May 1996, vol. 2, pp. 921–924.

[3] F. Keyrouz, Y. Naous, and K. Diepold, "A new method for binaural 3-D localization based on HRTFs," in *Proc. ICASSP06*, Toulouse, France, May 2006, vol. 5, pp. 341–344.

[4] M. Takimoto, T. Nishino, and K. Takeda, "Estimation of a talker and listener's positions in a car using binaural signals," in *Proc. the 4th Joint Meeting ASA and ASJ (ASA/ASJ06)*, Honolulu, Hawaii, USA, November 2006, p. 3216.

[5] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proc. ICASSP04*, Montreal, Canada, May 2004, vol. 2, pp. 817–820.

[6] B. Raj, M. V. S. Shashanka, and P. Smaragdis, "Latent direchlet decomposition for single channel speaker separation," in *Proc. ICASSP06*, Toulouse, France, May 2006, vol. 5, pp. 821–824.

[7] A. Fuchs, C. Feldbauer, and M. Stark, "Monaural sound localization," in *Proc. Interspeech 2011*, Florence, Italy, August 2011, pp. 2521–2524.

[8] R. Kliper, H. Kayser, D. Weinshall, I. Nelken, and J. Anemuller, "Monaural azimuth localization using spectral dynamics of speech," in *Proc. Interspeech 2011*, Florence, Italy, August 2011, pp. 33–36.

[9] R. Takashima, T. Takiguchi, and Y. Ariki, "HMM-based separation of acoustic transfer function for single-channel sound source localization," in *Proc. ICASSP10*, Dallas, Texas, USA, March 2010, pp. 2696–2699.

[10] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, Cambridge, MA, 2006.

[11] A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," *Journal Statistics and Computing*, vol. 14–3, pp. 199–222, August 2004.

[12] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy, "SVM and kernel methods matlab toolbox," Perception Systemes et Information, INSA de Rouen, Rouen, France, 2005.

[13] W. L. Chao, J. Z. Liu, and J. J. Ding, "Facial age estimation based on label-sensitive learning and age-specific local regression," in *Proc. ICASSP 2012*, 2012, pp. 1941–1944.

[14] S. Nakamura, "Acoustic sound database collected for hands-free speech recognition and sound scene understanding," in *Proc. International Workshop on Hands-Free Speech Communication (HSC01)*, Kyoto, Japan, April 2001, pp. 43–46.