Article:
Chen, Jinhui, Takashima, Ryoichi, Guo, Xingchen et al. (4 more authors) (2021) Multimodal Fusion for Indoor Sound Source Localization. Pattern recognition. 107906. ISSN 0031-3203

https://doi.org/10.1016/j.patcog.2021.107906

Multimodal Fusion for Indoor Sound Source Localization

Jinhui Chen[a], Ryoichi Takashima[b], Xingchen Guo[c,*], Zhihong Zhang[d], Xuexin Xu[d], Tetsuya Takiguchi[b], Edwin R. Hancock[e]

[a] *Prefectural University of Hiroshima, Hiroshima, Japan*
[b] *Kobe University, Kobe, Japan*
[c] *Xian Jiaotong University, Xian, China*
[d] *Xiamen University, Xiamen, China*
[e] *University of York, York, UK*

Abstract

To identify the localization of indoor sound source, especially when attempted using only a single microphone, it is a challenging problem to machine learning. To address these issues, this paper presents a distinct novel solution based on fusing visual and acoustic models. Therefore, we propose two novel approaches. First, to estimate orientation of vocal object in a stable manner, we employ the visual approach as estimation model, where we develop a robust image feature representation method that adopts Fourier analysis to efficiently extract polar descriptors. Second the distance information is estimated by calculating the signal difference between transmit receive ends. To implement these, we use phoneme-level hidden Markov models (HMMs) extracted from clean speech sound, to estimate the acoustic transfer function (ATF), which can capture the speech signal as a network of phoneme HMMs. And using the separated frame sequences of the ATF, we can indicate the signal difference between two positions, which can be used to estimate the distance of sound source. Experimental results show that the proposed method can simultaneously extract the sound source parameters of direction and distance, and thus improves the verification task of sound source localization.

*Corresponding author: Xingchen Guo
Email address: 80943667@qq.com (Xingchen Guo)

Keywords: Sound source localization, acoustic transfer function, HMM, polar HOG, SVM.

1. Introduction

The goal of this study is to automatically identify the position information of a dominant sound source which may alternate frequently among multiple candidate positions in the environment of an enclosed room. Methods for our task have to estimate the direction information of a sound source and its distance information. To simultaneously implement these, it is difficult, but this is an important capability in various applications in a number of applications, including teleconferencing scenarios [1], disease detection [2], speaker verification [3] and human-robot interaction (HRI) [3, 4]. Most of these applications require real-time processing of the signals. Furthermore, the estimation of sound source location is frequently used in the subsequent processing stages, such as sound source separation [5], sound source classification [6] and automatic speech recognition [7].

For the tasks of verifying the localization information of indoor sound source in the artificial intelligence system, there are a plethora of related studies described in the literature, but these mainly rely only on audio information and use the difference of arrival times for source localization using microphone arrays [3, 8]. However, microphone-array based systems depend on bulky equipment and are often computationally expensive, making them almost impracticable for real-time speech processing applications. Meanwhile, it is inevitable to range the microphones in a suitable length: when the microphones are spaced too closely together so that they all record essentially the same sound because of the Interaural Time Difference (ITD) near zero, making it extremely difficult to estimate the orientation [9, 10].

In this context, the ability to localize sound using a single microphone has emerged as an interesting low-complexity sound processing problem. In fact, sound source verification with the technique of single-channel microphone could

potentially be applied to a wider range of devices, especially small low-power devices with limited computational resources. Examples include wearable devices and smart phones, both of which have important potential commercial applications. Moreover, they also have interesting untapped potential in disaster relief tasks. For instance mobile sensors can be used to localize buried earthquake localize victims under rubble by following their voice.

However, since they have to simultaneously extract separate parameters for orientation and distance features, the accuracies of these existing methods are therefore still quite low. Alternative existing verification techniques for sound source location with a single microphone mainly use learning-based mapping procedures, accompanied with the use of external pinnae and/or inner-ear canals. Studies focusing on the techniques for monaural sound source localization are also being carried out [11, 12]. In these studies, the information obtained from the external ear, such as head-related transfer functions (HRTFs), is used to localize the sound source. For example, the work presented in [13] can only locate the types of sources for which it is trained, and as a result its performance might be affected by extraneous sound sources. One way around this problem is presented in [14], where a neural network is trained with recorded sound sources with known locations and different array positions. However, its distance estimation suffers from poor performance. This is because these methods [15, 13, 14] mainly focus on the estimation of direction (angle direction) between the sound source and the receiver (microphone). To the best of our knowledge, there has not been any work in the past aimed at the sound source/microphone absolute distance estimation from received single microphone signals.

Most of the work so far on localization of sound source relies solely on the acoustic modality. While audio-only-based techniques might present promising results, leveraging from visual information is often beneficial when a video camera is available. In the work reported in this paper, we adopt a visual model to implement the source direction estimation and then we estimate the distance of the sound source using an acoustic model, respectively. Moreover, both of these two feature modalities can be trained by a SVM. Our direction estimation

model is based on the convex imaging theory of a camera (see Fig. 1). To accurately estimate the orientation angle of a sound source, we propose a robust image feature extraction method based on Fourier analysis. This allows us to densely and efficiently calculate descriptor Fourier coefficients on a pixel-by-pixel basis and thus extract histograms of oriented gradients [16] (HOG)-type image polar features. In the Fourier domain, the polar data can be conveniently computed, since the rotation calculation is just a multiplication with a complex basis. Moreover, the polar features are sufficiently stable to accommodate the shaking of the bounding box, guaranteeing correct direction estimation.

Generally, a speech sound signal in a room environment can be represented as the convolution of clean speech sound and the acoustic transfer function (ATF). However, only AFT is a useful information for estimating the distance. Therefore, here we develop a novel method *i.e.*., phoneme Hidden Markov Models (HMMs) [17], to separate the ATF. However, to estimate the ATF, the separation approaches implemented by HMM, require texts of the speech utterances. In [18], although the utterance texts for the adaptation data are given, the case for unsupervised on-line adaptation was not discussed. In the studies of unsupervised adaptation for speech recognition, it is able to obtain the utterance text by using the word recognition approaches [18, 19]. However, in the case of verification tasks for sound source localization, we require the utterance text of the test data without any dictionaries or knowledge of the language. Therefore, the alternative method is used for the proposed method that can identify the information of sound source localization in phoneme level. Our solution uses a classification system to recognize the phoneme to replace the conventional text information for the ATF estimation. Since the cepstral parameters are the effective representation model for preserving useful clean speech sound information, the estimation approaches are performed in the cepstral domain adopting the method based on maximum likelihood (ML).

Experiments demonstrate the effectiveness of our methods for the sound source verification task using a single-channel microphone. Our study makes the following distinct contributions:

- Accurately, we explore a novel solution to accurately identify localization information of sound source, which fuses visual and acoustic models based on a single-microphone into a multimodal framework. Since the task of single-microphone voice source localization is one of the most challenging scenarios in the area of speech signal processing, our solution is therefore likely to lead to further related research.

- We develop an HMM-based method for separation of the ATF to describe clean speech sound. This leads to the accurate indication of the temporal phonetic changes of clean speech sound in a single channel.

- We propose a new Fourier domain method for fast implementation of the HOG-type polar feature descriptor. The proposed method simultaneously has rotation-invariant capabilities and preserves the discriminative power of extracted features.

The remainder of this paper is organized as follows. In Section 2, we review the literature on related work, and discuss the relationship between our proposed model and a number of alternative methods. Section 3 gives overview of proposed method. Sections 4 and 5 respectively illustrate the proposed acoustic model for estimating sound source orientation and sound source distance with the proposed visual model. This is followed by an experimental evaluation in Section 6. Finally, conclusions are presented in Section 7.

2. Related Works

To date, most of the work on the localization of sound sources relies solely on the acoustic modality. The task is usually considered as being composed of two parts, namely (1) determining the direction of arrival (DOA) of the source and (2) determining the distance of the source to the microphone.

2.1. Direction of arrival (DOA) estimation

In signal processing tasks, DOA can be estimated using a variety of approaches such as source clustering through time and tracking techniques using

Kalman filters [20, 21] or particle filtering [22]. The implementation of these techniques depends on the number of microphones. For a single-channel, it is common to make use of the time-difference-of-arrival (TDOA) between a pair of sensors or microphones. In the past, the most popular way to estimate the TDOA is based on calculating a cross correlation vector [23]. To improve the performance of this approach, Fourier domain methods are used to efficiently compute the generalized cross-correlation (GCC) [24]. Problems are also encountered with GCC-based approaches, because Dirac delta functions appear in the correlation vector in the case of high correlation and the Fourier transform of a Dirac delta function spans the complete frequency domain. One way to enhance this approach-based approaches is to use the phase transform[25]. However, as mentioned in the previous section, these approaches are very sensitive to equipment reverberations and other noise sources, and GCC cannot be used with only one microphone.

In this study, our solution is to use visual processing tools to calculate the direction of the sound source. The proposed image feature used in our work is reminiscent of Dalal *et al.*'s histogram of oriented gradients (HOG) [16]. The HOG feature was first proposed to represent objects in images using the distribution of gradient magnitude and orientations over spatially distributed regions [26]. It has been widely acknowledged as one of the best features to capture edge or local shape information of the objects. More recently improved HOG-type descriptors have been developed including histograms of radial gradient [27] and Oi-HOG [28] that leverage radial or polar gradient transforms to achieve rotational invariance. Although the aforementioned feature representations have shown impressive levels of success on a variety of visual tasks, they are highly dependent on the relative position with respect to the center of the local patch. Furthermore, they require us to calculate an auxiliary space (unit vectors of tangent and radius directions) in polar coordinates. These factors limit the performance in some applications.

2.2. Distance estimation

After the direction-of-arrival is computed, then the distance to the source must be estimated to achieve sound source location . A widely used approach relies only on audio information and utilizes time delay cues for localization by microphone arrays [3, 8]. This method splits the microphone array into pairs and estimates the time difference of arrival (TDOA) for different microphone pairs [29]. This can produce significant distance estimation errors. However, microphone-array based systems depend on bulky equipment and are often computationally expensive, making them virtually impracticable for real-time speech processing applications.

Recently, there has been work aimed at using monophonic signals. These include a neural network based approach [14], where distance is not estimated directly. Instead, it is a byproduct of estimating the Cartesian coordinates of the source using just one microphone. Additional existing monophonic techniques [15, 13] mainly focus on the estimation of direction (angle detection) between the sound source and the receiver (microphone). Lu *et al.* [30, 31] have proposed a binaural distance estimator for the dynamic case in which the receiver is moving. Georganti *et al.* [32] have proposed a novel detector based on extraction of statistical measures from the single-channel microphone in combination with the classification-based Gaussian mixture model. Smaragdis and Boufounos [29] have employed an expectation maximization algorithm for learning the amplitude and phase differences of cross-spectra in order to determine the position of a sound source using two microphones. This method was later improved by Vesa [33] in order to account for different positions that have the same azimuth angle. The method makes use of the magnitude-squared coherence, which is a frequency-dependent feature. The feature is used in addition white noise in the training of a Gaussian Maximum Likelihood scheme for distance estimation.

To the best of our knowledge, there has been no prior work aimed at sound source-to-microphone absolute distance detection from received monophonic speech signals. In one-microphone tasks, there are a number of problems to solve. These include a) changes in the talker characteristics, b) the speaker

position and c) the room environment. As discussed in the previous section, the accurate estimation of the ATF from the observed speech is important for these scenarios. Therefore, we focus on the ATF to estimate the distance of the sound signal source. Since, an HMM can describe the features of all phonemes more accurately, the proposed method adopts a phoneme HMM to accurately separate the ATF.

3. The Overview of Proposed Method

As explained above, our solution to identify the sound source position information fuses together both visual and acoustic models to improve accuracy. Both the visual and the acoustic models are trained by SVM [34].

3.1. The Proposed Visual Model Overview

As illustrated in Fig. 1, we adopt the convex imaging theory of camera to estimate the orientation of the sound source, since this offers one of the most succinct and effective solutions available. In Fig. 1 , we consider the central point $r(x, y, z)$ of the object as the observed point for orientation calculation. The imaged point corresponding to this point on the image plane of the camera is $r'(x', y', z')$ (the units of x', y' and z' are pixel, pixel and cm, respectively) in the bounding box of the sound source object (that is detected by the proposed object detection model). According to the convex projection theorem, the orientation angle α of the object is calculated as follows,

$$\alpha = \arctan\frac{(x' - x'_0)\Delta d}{z'}, \tag{1}$$

where z', $c(x'_0, \cdot, \cdot)$, and Δd denote image distance, the central point on the image plane, and the pixel pitch respectively.

3.2. The Proposed Acoustic Model Overview

Fig. 2 gives an overview of the proposed acoustic model. The reverberant speech sound signal in the room environment can be represented by the convolution of clean speech sound frame sequences and the ATF. We can adopt the ATF

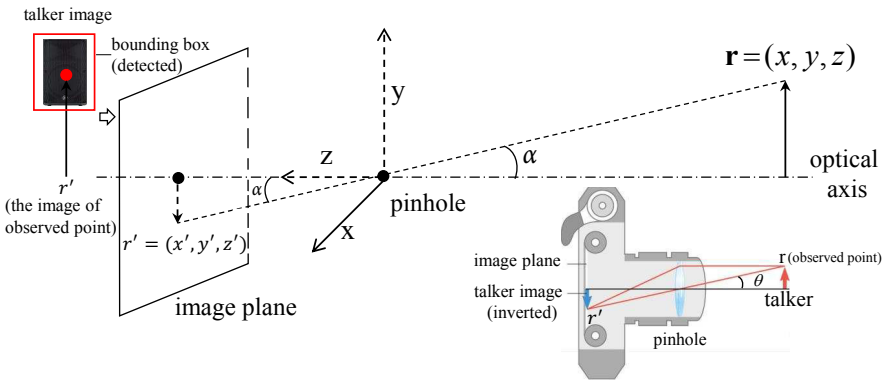


Figure 1: Illustration of sound source directional estimation.

to estimate the sound source distance. The training stage of proposed acoustic model includes three approaches: a) the data set $O_{train}^{(\theta)}$ of reverberant speech sound uttered from each distance position θ are firstly recorded as preparations. b) We adopt a clean speech sound model that is trained in advance by using a clean speech sound database, to estimate $\hat{H}_{train}^{(\theta)}$ from $O_{train}^{(\theta)}$, where $\hat{H}_{train}^{(\theta)}$ and $O_{train}^{(\theta)}$ indicate frame sequences of ATF and reverberant speech sound, respectively. c) Frame sequences of the estimated ATF $\hat{H}_{train}^{(\theta)}$ are fed into the support vector machine (SVM) to train the classification model to estimate the distance of the sound source position θ. In the test stage, the ATF $\hat{H}_{test}^{(\theta)}$ is estimated from input data $O_{test}^{(\theta)}$ (any utterance) in the same way as the training procedures. The distance of sound source position $\hat{\theta}$ is estimated by the trained classification model.

4. Estimation of Source Orientation with the Proposed Visual Model

As already discussed Subsection 3.1, direction estimation is based on the convex imaging theory, using a camera. In this approach, the parameters of the pixel pitch Δd and image distance z' can be deduced from the camera imaging system. However, we have to obtain the observed position r' of the sound source

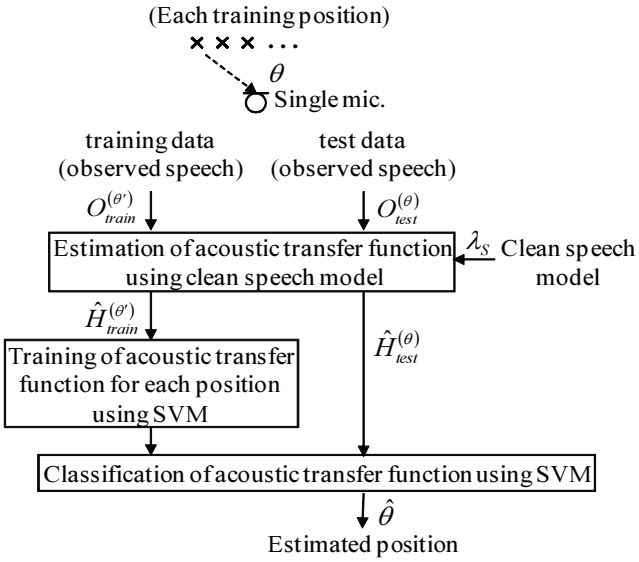


Figure 2: The proposed acoustic model overview, where θ denotes the distance of the sound source position.

object in the image. The orientation estimation is therefore dependent on estimating the range of the source object from the available images. Consequently, we require a robust object detection tool to bound the box containing the source object in the available images. This object detection tool should be capable of capturing the invariant features so that the object position can be obtained in a stable manner.

In our earlier work [28], we have systematically demonstrated the advantages and in particular the rotation-invariant capabilities of polar descriptors in local image feature representation. Polar data can either be represented in a Cartesian basis $\eta = [\alpha, \beta]^T \in \mathbb{R}^2$ or in a radial basis $[r, \varphi]$ (r and φ are the norm of η and the angle of η, respectively). The Cartesian approaches make use of the tangent directions in polar coordinates as the reference directions. They then generate rotation or reversal-invariant descriptors by normalizing the contents of multiple concentric spatial bins [28, 35]. However, this solution is dependent on a) auxiliary calculations of the tangent or radius vectors and b) the relative

position to a selected center in the local feature patch of the images. These two elements of the calculation provide a computational bottleneck and as a result there is a relatively limited potential for developing "fast" Cartesian features.

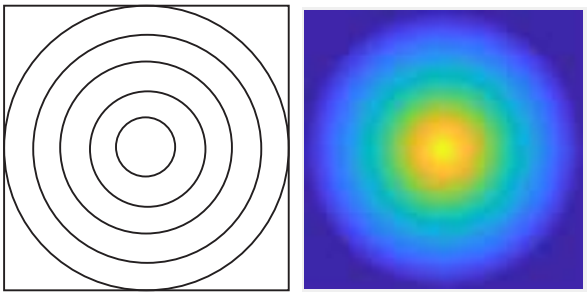
Since we require the computations of the image model to be synchronized with those of the acoustic model, the processing speed plays an important role in the choice of methods for our task. We therefore use a radial rather than a Cartesian basis to extract the polar HOG descriptor. There are two reasons for this. Firstly the polar basis can be used conveniently in conjunction with Fourier domain analysis to improve the HOG feature extraction efficiency. Secondly, the Fourier-based approach is naturally invariant to rotations.

A standard HOG feature is calculated in three procedures: a) binning the gradient orientation, b) spatial aggregation, and c) normalization [16]. In this paper, the orientation quantization for gradients is created using an orientation distribution function to make the image gradient for one pixel $p(x, y)$ be oriented to bin φ.

To calculate polar HOG descriptor η, we require a local patch function g to capture the structural information in the neighborhood surrounding a pixel. The local patch function can be implemented by using the Gaussian kernel. We calculate the local polar HOG descriptors at position (x, y) by collecting all gradient magnitudes within the local patch function g to calculate the orientation φ (which is similar to the procedure of spatial aggregation in HOG [16]). The polar HOG descriptor η at each pixel is computed as follows,

$$\eta(p, \varphi) = \int \|d(t)\| \, \delta(\hat{d}(p)) g(p - t) dt, \tag{2}$$

where φ=atan2(y, x), $d(\cdot)$ is the gradient, and $\hat{d}(\cdot) :=$ is a function to get the angle of $\frac{d(\cdot)}{\|d(\cdot)\|}$, respectively. In addition, the delta function (Dirac) $\delta(\cdot)$ is used to remove those gradients without orientation φ. $g(\cdot)$ is a Gaussian function, which is used to implement spatial aggregation by evaluating the feature contributing to orientation φ at point p. A polar function can be expanded linearly in terms of Fourier basis functions. Therefore in the Fourier domain, Eq. 2 can be



(a) Annular spatial cells **(b)** Polar HOG

Figure 3: Illustration of annular spatial cells and a polar gradient template image over the annular spatial patch.

represented as,

$$\eta(p,\varphi) = \frac{1}{2\pi}\sum_{m=-\infty}^{+\infty} n(p)e^{-im\varphi}. \tag{3}$$

By considering the Fourier expansion of the delta function $\delta(x) = \frac{1}{2\pi}\sum \overline{e^{-mx\varphi}}e^{-im\varphi}$, we can compute the descriptor as follows,

$$\eta(p,\varphi) = \frac{1}{2\pi}\sum_{m} \overline{\|d(p)\|\, e^{-ikm\varphi}} * g(p)e^{-im\varphi}, \tag{4}$$

where m is an integer, $k = \hat{d}(p)$. To obtain the rotation-invariant descriptors, we denote $t(\cdot) = \frac{1}{2\pi}\overline{\|d(p)\|\, e^{-ikm\varphi}} * g(p)e^{-im\varphi}$ and we can describe the rotated descriptor as follows,

$$\begin{aligned} R_{\Delta\varphi}\eta &= \sum_{m} \overline{\|d(p)\|\, e^{-ikm(\varphi-\Delta\varphi)}} * g(p)e^{-im(\varphi-\Delta\varphi)} \\ &= \sum_{m} e^{i(-km+m)\Delta\varphi} * t(\cdot). \end{aligned} \tag{5}$$

The rotation-invariant should follow the condition $\eta = R_{\Delta\varphi}\eta$ where R is a rotation matrix by a given angle $\Delta\varphi$, *i.e.*, $k = 1$.

Based on the above, we can densely compute Fourier representations of the candidate descriptors at each pixel point η from Eq. 4. For fast descriptor processing, we implement the calculation by decomposing it into two parts namely, $\eta_1(p,m) = \|d(p)\|\, e^{-ikm\varphi}$ and $\eta_2(p,m) = g(p)e^{-im\varphi}$. In our experiments, we

Algorithm 1 Fast Polar Feature Extraction in Fourier Space

Require:

A set of image data $\mathbf{I}$ over the feature local patch;

Ensure:

Polar HOG feature $\mathbf{T}(\mathbf{I})$;

Compute the gradient of image $\mathbf{I}$: $D := \nabla\mathbf{I}$;

Initialization: Set feature descriptor set $\mathbf{T} = \varnothing$;

Subdivide the local patch into annular spatial cells, as shown in Fig. 3(a);

for m=0:6 **do**

Calculate part descriptor basis in Fourier space for each pixel $p(x, y) \in \mathbf{I}$:

$$\begin{aligned}
&\varphi = \text{atan2}(y, x),\\
&k = \text{atan2}\frac{d(p)}{\|d(p)\|},\\
&\eta_1(p, m) = \|d(p)\| \, e^{-ikm\varphi},\\
&\eta_2(p, m) = g(p)e^{-im\varphi};
\end{aligned}$$

Normalize the descriptor

$$\eta_1(p, m) = \frac{\|d(p)\| \, e^{-ikm\varphi}}{\sqrt{\|D(\mathbf{I})\|^2 \bigotimes K}};$$

end for

for all η_1 **do**

if k=1 **then**

Calculate polar HOG descriptor in each cell:

$$\eta(p, m) = \eta_1(p, m) * \eta_2(p, m);$$

Add polar descriptors η to $\mathbf{T}$, according to their polar gradient orientations;

end if

end for

Output $\mathbf{T}$;

find it difficult to implement the normalization of histograms in Fourier space. We therefore normalize the descriptor before sorting them into bins. Since η_2 uses a Gaussian kernel to implement the spatial aggregation[16, 35], we require only a local normalization calculation. Following [36, 35], we adopt a convolution operator to implement the local spatial normalization as follows,

$$\eta_1(p,m) = \frac{\|d(p)\|\, e^{-ikm\varphi}}{\sqrt{\|D(\mathbf{I})\|^2 \bigotimes K}}, \tag{6}$$

where $D := \nabla \mathbf{I}$ is the gradient of the image $\mathbf{I}$ ($d(p) \in D(\mathbf{I})$) and K is a smoothing convolution kernel. The procedure for extracting polar HOG descriptor in Fourier space is described in Algorithm 1.

We use Algorithm 1 to densely calculate the patch of feature templates in an image. In this paper, we set the maximum patch size of these templates as 100 $\times$ 100 pixels (see an example of descriptors generated by Algorithm 1 in Fig. 3(b)). In addition, we allow different aspect ratios for each template patch (*i.e.* the ratio of width to height). In the Fourier domain, we can densely calculate polar gradients on local patches. Furthermore, since the parameter of the polar descriptors can be conveniently calculated by projecting them into Fourier space, we can efficiently extract descriptors to meet the speed requirements necessary to synchronize with the acoustic model for verifying the sound source localization information.

We use the proposed feature to detect the object of sound source. The detector window is tiled with the above feature patch in which polar vectors are extracted. The combined vectors are fed to the SVM for object/non-object classification. The detection window is scanned across the image at all positions and scales, and conventional non-maximum suppression is run on the output pyramid to detect the sounding object. Finally, the orientation can be calculated in accordance with procedures of Subsection 3.1.

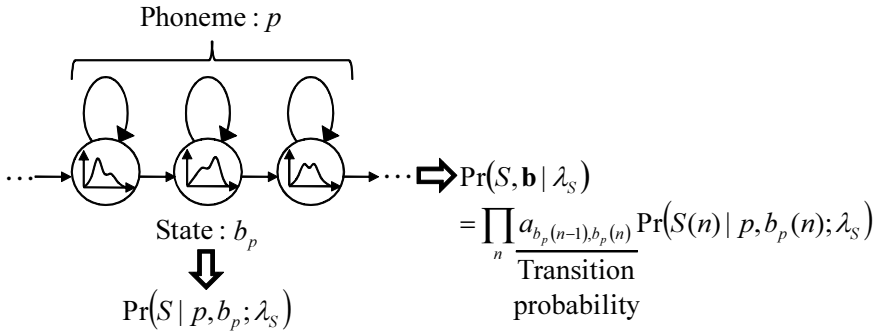


Figure 4: Clean speech model expressed by an HMM, where S denotes the clean speech sound signal, λ denotes the parameter set of HMM, and $a_{b_p(n-1),b_p(n)}$ indicates the transition probability.

5. Estimation of Source Distance with the Proposed Acoustic Model

In this study, we adopt clean speech sound HMMs to estimate the frame sequence of the ATF $\hat{H}^{(\theta)}$. For following easily, an example for illustrate the HMM model of a clean speech sound is shown in Fig. 4. In the case of HMM representation, the speech is represented by a state transition model, and the posterior probability estimation is implemented with the Gaussian mixture model (GMM) [37] for each state. Therefore, we can calculate the likelihood as the product of the transition probability $a_{b_p(n-1),b_p(n)}$ and the posterior probability for each state of the frame n. Namely, we can express the clean speech sound as a network of phoneme HMMs. In this clean speech sound HMM model, since the more detailed clean speech sound information, such as the temporal phonetic changes can be represented by the individual phoneme HMMs, we therefore can preserve the discriminative structure parameters of the ATF for the distance estimation.

Fig. 5 shows the overview of the proposed distance estimation procedures. Our method is based on the ATF that is represented by the HMMs of clean speech sound. Since the HMMs of clean speech sound are trained for each

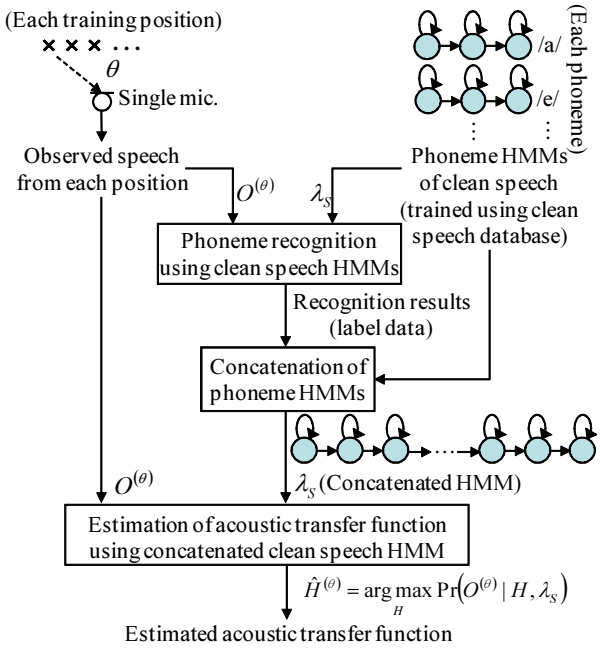


Figure 5: Estimation of the ATF using phoneme HMMs of clean speech sound.

individual phoneme, to construct the network of phoneme HMMs, we require the texts of user utterances. Consequently, we firstly use each phoneme HMM derived from the clean speech sound data to recognize the phoneme sequence of the reverberant speech sound data. Then, we adopt the recognized results (1-best hypothesis) to concatenate the network of phoneme HMMs. Finally, the frame sequence of the ATF $\hat{H}^{(\theta)}$ is estimated from the reverberant speech sound $O^{(\theta)}$ by the maximum likelihood estimation based on the concatenated HMM.

5.1. Cepstrum Representation of Reverberant Speech

Generally, the reverberant speech sound signal $O(t)$ in a room environment is represented as the convolution of clean speech sound and the ATF $O(t) = \sum_{l=0}^{L-1} S(t-l)H(l)$, where $S(t)$, $H(l)$, and L denote clean speech sound signal, ATF (room impulse response) from the sound source to the microphone and the

length of the ATF, respectively.

Recently, some related works [38, 39, 40] have focused on the modelling of reverberant speech sound in the short-term Fourier transform (STFT) domain, presenting solutions to tasks of speech recognition and speech dereverberation. Specifically, each frequency bin of the reverberant speech sound can be represented by the ATF and the convolution of the clean speech sound frame sequences as follows,

$$O_{spc}(\tau; n) = \sum_{l'=0}^{L'-1} S_{spc}(\tau; n - l') \cdot H_{spc}(\tau; l'), \tag{7}$$

where O_{spc}, S_{spc} and H_{spc} respectively denote spectra for the reverberant speech sound signal, the clean speech sound signal and the ATF. τ is the index number of frequency bins of the short-term linear spectra in the n-th frame sound signal sequence, and L' express the length of the ATF in the STFT domain. However, the cost of such solution to estimate the frame sequence of the ATF is quite expensive [41]. Therefore, the estimated components of the ATF are too complex and it is difficult to deal with those parameters for this task.

To address the above problems, in this paper, we try to adopt a simpler model to represent the signal of reverberant speech sound. We consider that a linear spectra O_{spc} of **short term** can be approximately represented as $S_{spc}(\tau; n) \cdot H_{spc}(\tau; n)$. It has been widely known that spectra are not ideal model for sound signal feature representation, yet the cepstral-based feature can preserve an effective representation and discriminative information in tasks of speech recognition. Therefore, we require use the cepstrum to replace the spectrum to estimate the ATF. The cepstral representation for the reverberant speech sound signal is defined as follows, *i.e.*, the inverse Fourier transform of the log spectrum,

$$O_{cep}(d; n) \approx S_{cep}(d; n) + H_{cep}(d; n), \tag{8}$$

where the meaning of symbols O_{cep}, S_{cep} and H_{cep} is similar to the representation in Eq. 7. d denotes the dimensionality of a cepstrum. When O and S are

observed, we can simply obtain H as the following equation,

$$H_{cep}(d;n) \approx O_{cep}(d;n) - S_{cep}(d;n). \tag{9}$$

5.2. Maximum-Likelihood Parameter Estimation

As discussed in the last subsection, the Eq. 8 shows a possible solution to estimate the ATF, but we cannot observe the cepstrum of clean speech sound(S_{cep}) directly. We, therefore require an alternative solution to estimate the ATF H_{cep}. In this subsection, we propose an alternative approach for estimating the ATF, which is implemented by maximizing the likelihood of the training data from the sound source position. For simplicity in this subsection, the cepstral variables O_{cep}, S_{cep} and H_{cep} are written as O, S and H, respectively.

In order to estimate the frame sequence of the ATF in Eq. (9), we adopt the expectation maximization algorithm to maximize the likelihood of the observed sound as shown follows,

$$\hat{H} = \underset{H}{\operatorname{argmax}} \Pr(O|H, \lambda_S), \tag{10}$$

where λ_S indicates the parameter set of concatenated clean speech sound HMM and the subscript S denotes the index refer to the cepstral domain of the clean speech sound. As we known, the expectation maximization algorithm has two iterative procedures. In the first procedure, referred to as the expectation step, the aim is to calculate the following expected log-likelihood function,

$$\begin{aligned} & f(\hat{H}|H) \\ &= E[\log \Pr(O, p, b_p, c_p|\hat{H}, \lambda_S)|H, \lambda_S] \\ &= \sum_{p,b_p,c_p} \frac{\Pr(O, p, b_p, c_p|H, \lambda_S)}{\Pr(O|H, \lambda_S)} \cdot \log \Pr(O, p, b_p, c_p|\hat{H}, \lambda_S), \end{aligned} \tag{11}$$

where b_p denotes the unobserved state of the speech sound frame sequence, and c_p represents the unobserved mixture component labels corresponding to the phoneme p in the observation sequence O.

The joint probability of frame sequences O, b and c, which are being observed, can be expressed as follows,

$$\begin{aligned}&\Pr(O, p, b_p, c_p|\hat{H}, \lambda_S)\\&= \prod_n a_{b_p(n-1),b_p(n)} w_{b_p(n),c_p(n)} \cdot \Pr(O(n)|p, b_p(n), c_p(n); \hat{H}, \lambda_S),\end{aligned} \tag{12}$$

where a and w respectively denote the transition probability and mixture weight. Since we consider the ATF as subject to additive noise in the cepstral domain, the mean of the mixture k at the state j in the model λ_O is derived by adding the ATF, where λ_O indicates the parameter set of concatenated reverberant speech sound HMM and the subscript O denotes the index refer to the cepstral domain of the reverberant speech sound. Therefore, Eq. (12) can be written as,

$$\begin{aligned}&\Pr(O, p, b_p, c_p|\hat{H}, \lambda_S)\\&= \prod_n a_{b_p(n-1),b_p(n)} w_{b_p(n),c_p(n)} \cdot \mathcal{N}(O(n); \mu^{(S)}_{p,j,k} + \hat{H}(n), \Sigma^{(S)}_{p,j,k}),\end{aligned} \tag{13}$$

where $\mathcal{N}(\cdot)$ is the distribution of multivariate Gaussian. The following derivation of the expression for the expected log-likelihood is straightforward [42]:

$$\begin{aligned}&f(\hat{H}|H)\\&= \sum_{p,i,j,n} \Pr(O(n), p, b_p(n) = j, b_p(n-1) = i|H, \lambda_S) \log a_{p,i,j}\\&\quad + \sum_{p,j,k,n} \Pr(O(n), p, b_p(n) = j, c_p(n) = k|H, \lambda_S) \log w_{p,j,k}\\&\quad + \sum_{p,j,k,n} \Pr(O(n), p, b_p(n) = j, c_p(n) = k|H, \lambda_S)\\&\quad \cdot \log \mathcal{N}(O(n); \mu^{(S)}_{p,j,k} + \hat{H}(n), \Sigma^{(S)}_{p,j,k}),\end{aligned} \tag{14}$$

where μ and $\Sigma^{(S)}$ respectively represent the mean and the covariance in the concatenated HMM network of clean speech sound, and i denotes the state index number for previous sequences. Since these parameters can be calculated by using a clean speech sound dataset, it is able to train the classification framework based on these parameters for estimating the distance of sound source. Since the covariance $\Sigma^{(S)}$ is a diagonal matrix, the expression of the Eq. 14 just simply

calculates those terms involving the cepstrum ATF H as follows,

$$f(\hat{H}|H) = h \sum_{p,j,k,n} h_{p,j,k}(n) \sum_{d=1}^{D} \left\{ \frac{1}{2} \log(2\pi)^D \sigma_{p,j,k,d}^{(S)^2} + \phi \right\}, \tag{15}$$

where $h_{p,j,k}(\cdot) = \Pr(O(\cdot), p, j, k|H, \lambda_S)$, $\phi = \frac{(O(d;n)-\mu_{p,j,k,d}^{(S)}-\hat{H}(d;n))^2}{2\sigma_{p,j,k,d}^{(S)^2}}$. D denotes the dimension of $O(n)$. $\mu_{p,j,k,d}^{(S)}$ and $\sigma_{p,j,k,d}^{(S)^2}$ are the d-th element of its mean and the d-th diagonal element of its covariance matrix, respectively.

Based on the analyses above, the maximization procedure of our method for expectation maximization algorithm is to maximize the likelihood $Q(\hat{H}|H)$. The function for updating H can therefore be derived from the condition $\partial Q(\hat{H}|H)/\partial \hat{H} = 0$ as,

$$\hat{H}(d;n) = \frac{\sum_{p,j,k} h_{p,j,k}(n) \frac{O(d;n)-\mu_{p,j,k,d}^{(S)}}{\sigma_{p,j,k,d}^{(S)^2}}}{\sum_{p,j,k} \frac{h_{p,j,k}(n)}{\sigma_{p,j,k,d}^{(S)^2}}}. \tag{16}$$

When the calculation of ATF over all sound data of different distances has been executed by using the proposed approaches above, we can obtain the estimated ATF. In this study, the obtained signal feature represented by the transfer function in different distances is then fed into SVM for training.

6. EXPERIMENTS

The proposed method was carried out in both of simulated and real conditions for evaluations of ideal environment and real environments. The visual model for this task cannot be accurately synthesized in a simulated environment. Therefore, in our simulated experiments, we used the proposed acoustic model to estimate both distance and orientation for the sound source. In our real-world experiments on the other hand, we embed the visual model into the proposed method to test the method on representative applications involving sound source identification.

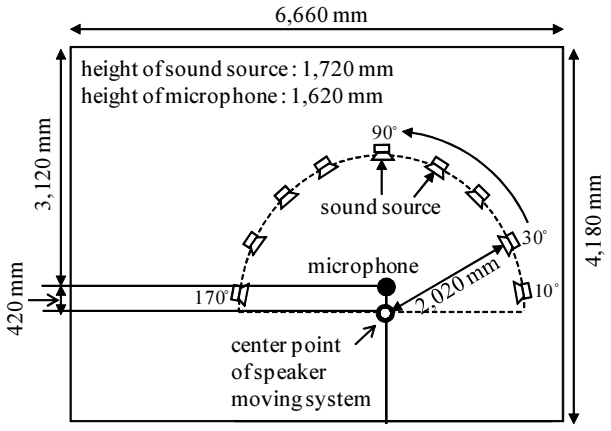


Figure 6: Experimental room environment for simulation.

6.1. Simulation-experiment Dataset and Implementation

Fig. 6 illustrates the environment setting for simulation experiments. The width and length of the simulation experimental room were 6.7 meters and 4.2 meters, respectively. The reverberation time in this room was 300 milliseconds. A loudspeaker acted as sound source (the height of that was 1,720 millimeters) was located on a semicircle with the 2,020-millimeter radius. A single microphone as the input device (the height of that was 1,620 millimeters) was set 420 millimeters from the circle center. The loudspeaker was set to face the input device. Therefore, the distances between the microphone and each position of sound source ranged from 1,600 to 1,900 millimeters.

In the simulated experiments, the reverberant speech sound from every position was simulated by linearly computing the convolution of clean speech sound and the ATF (impulse response). The impulse response data were estimated by clean speech sound from the RWCP database of real acoustical environments [43, 44]. The estimated ATF was then fed into SVM for training to obtain the classification model of sound source localization.

We adopt the ATR Japanese speech database as sound data. The speech

Table 1: The MSE results for the ATF separated by different models.

	The observed speech	GMM	HMM (1-best)	HMM (CT)
MSE	9485.97	2264.33	2096.14	1968.36

sound signal was sampled in frequency of 12 kHz, with the 32-millisecond Hamming window, and the frame shift was 8 milliseconds. A 16-order mel-frequency cepstral coefficients (MFCCs) [45] was used as the features. There are totally 54 phonemes in Japanese. The proposed HMM model for every phoneme was a simple left-right model, which has three self-transition states. Each state contains 32 Gaussian mixture components.

We used 2,620 words to train the clean speech sound HMM (speaker-dependent model), and used 10, 20, 30, 40 and 50 words to respectively train the ATF and to evaluate their performances for one location. The number of test data was 1,000 words per person section for each location. There were five randomly selected persons. The training data (for clean speech sound model O and the ATF H) and testing data were spoken by the same person, but the speech contents of that were different.

We also set different conditions of sound source positions for evaluations. We trained proposed model and tested its performances in three positions set of 30°, 90° and 130 °, five positions set of 10°, 50°, ..., 170°, seven positions set of 30°, 50°, ..., 150° and nine positions set of 10°, 30°, ..., 170°.

6.2. Results of the Simulated Reverberant Environment

We had tried a large number of candidate methods to implement the ATF estimation and the following methods obtained acceptable results. We therefore adopted these models as comparison methods.

The observed speech This comparison method was implemented by directly classifying the observed speech without separating the ATF for verifying the sound source localization information.

GMM Similar to the proposed HMM-based approach, this comparison method used a clean speech GMM to estimate the ATF. In a manner different from HMM-based approaches, each stage of phoneme model has 64 Gaussian mixture components.

HMM (1-best hypothesis) The HMMs network of phoneme in the proposed model are concatenated using those phonemes with best recognition results (1-best hypothesis).

HMM (correct transcription) The HMMs network of phoneme in the proposed model are concatenated by using the correct transcription.

To make fair comparison, we used the same learning framework, the same data and the same experimental conditions for these methods.

Table 1 shows the mean square error (MSE) for the ATF separated by different models, where GMM, HMM (1-best) and HMM (CT) respectively denote the models based on clean speech sound GMM, clean speech sound HMMs with the 1-best hypothesis, and HMMs with the correct transcription. We can calculate the MSE as follows,

$$\mathrm{MSE} = \frac{1}{N}\sum_{n,d}(H_{\mathrm{true}}(d;n) - \hat{H}(d;n))^2, \tag{17}$$

where $\hat{H}$ is the estimated ATF. H_{true} denotes the calculation result of the true clean speech sound data from the Eq. 9, referred to as the ground truth of the ATF, d denotes the d-th dimension of a cepstrum in the n-th frame sound signal sequence, and N express the length of the cepstrum representation. As results shown in the table, the MSE of the proposed approach was smaller than comparison methods. This means that the proposed method can estimate the ATF more accurately than other methods.

In our method, the acoustic transfer function is separated from the training speech uttered by the same people from the same position in the room as those for testing. The related evaluations are shown in Fig. 7. The 32-order mel spectrum was obtained by computing the inverse cosine transform of the 32-order MFCCs, where the estimated 16-order MFCCs were extended to 32-order MFCCs using zero padding. Then, the mel spectra had normalized energies

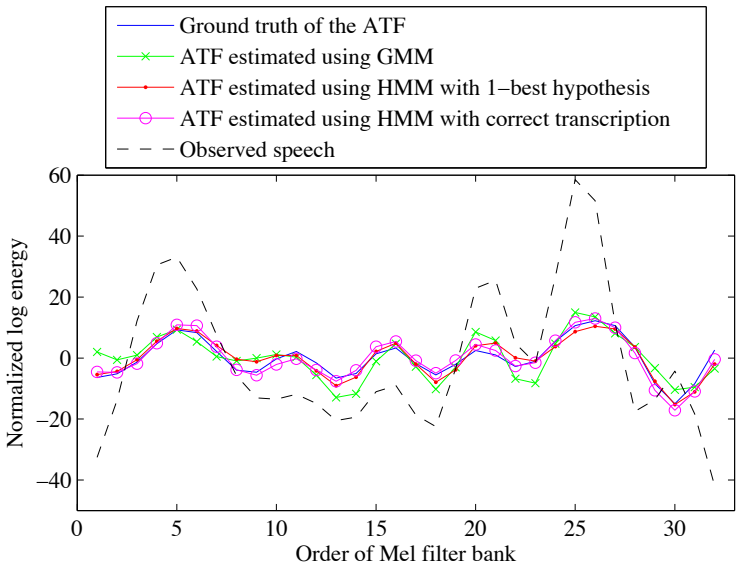


Figure 7: Mel spectra of different ATF models compared with that of the ground truth.

because the 0th dimension of the MFCCs (which is equivalent to the energy of the mel spectrum and is not discussed in this study) was also padded with zeros. This figure also shows that the clean speech HMMs suppressed the influence of clean speech more effectively and estimated the acoustic transfer function more correctly than the existing methods.

Table 2 shows a comparison of the different methods for using 10-, 20-, 30-, 40- and 50-word training data, where the average accuracy was calculated in three-position set. The parenthetic numbers show the differences from the accuracy of the HMMs with 1-best hypothesis. Table 2 also shows that as the number of training data decreased, the accuracy of the proposed method became much larger than other methods. These results mean that the proposed method will obtain better performances, especially in the cases of small-size training data. Because in the classification framework based on the observed speech, when the number of training data decreased, the classification boundaries were biased by

Table 2: Localization accuracies [%] (3 positions) of compared different methods for using 10-, 20-, 30-, 40- and 50-word training data.

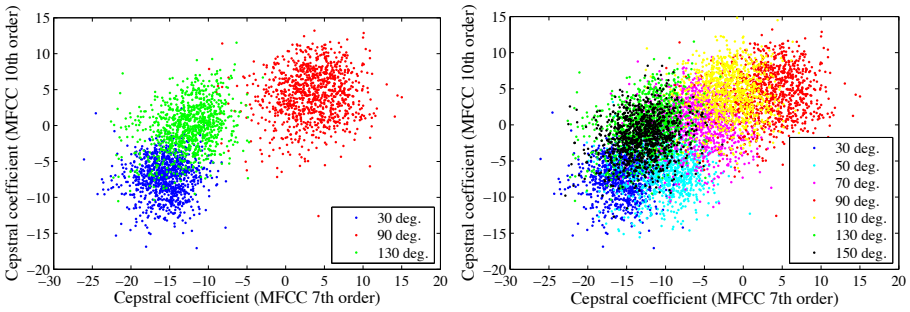
Number of training data (words)	50	40	30	20	10
HMM (1-best hypothesis)	82.9	82.4	82.3	80.6	79.9
HMM (correct transcription)	83.9 (0.9)	83.3 (0.9)	83.0 (0.8)	81.0 (0.4)	79.5 (-0.4)
GMM	80.5 (-2.4)	80.2 (-2.2)	79.2 (-3.1)	76.9 (-3.6)	75.1 (-4.8)
Observed speech	53.2 (-29.7)	50.2 (-32.2)	46.6 (-35.7)	40.7 (-39.9)	35.7 (-44.2)

Table 3: Localization accuracies [%] (50 words) of compared methods for each number of positions.

Number of positions	3	5	7	9
HMM (1-best hypothesis)	82.9	61.4	56.0	46.6
HMM (correct transcription)	83.9 (0.9)	62.0 (0.7)	58.0 (1.9)	48.0 (1.5)
GMM	80.5 (-2.4)	57.0 (-4.4)	53.6 (-2.5)	44.0 (-2.5)
Observed speech	53.2 (-29.7)	27.0 (-34.4)	30.3 (-25.7)	22.5 (-24.1)

the utterance contents of the training data, which will cause overfitting. Similar situation also exists in the GMM-based frameworks, because the clean speech sound components are not completely removed from the observed signal in those frameworks.

Table 3 indicates comparisons of the different methods for identifying sound source localization of 3, 5, 7, and 9 positions, where the number of training data was 50 words. The parenthetic numbers show the differences from the accuracy of the HMMs with 1-best hypothesis. The observed speech includes information of the acoustic transfer function and clean speech sound. The clean speech sound is not useful information for sound source localization, which also requires addi-



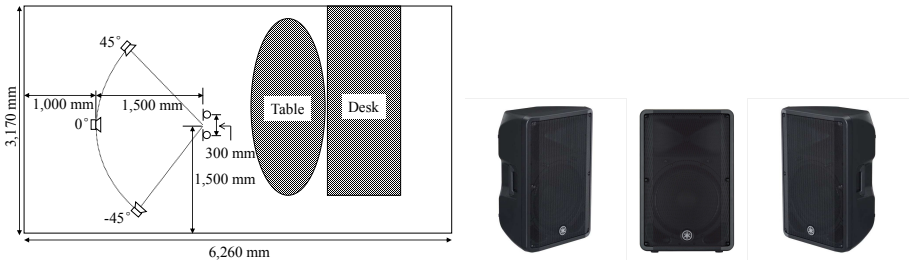
(a) Three positions **(b)** Seven positions

Figure 8: Mean ATF values.

tion calculation cost. Since our methods separate the acoustic transfer function from the observed speech signal, the proposed method showed higher accuracies than the method adopted the observed speech. In the comparison between the proposed method (HMMs with 1-best hypothesis) and clean speech sound GMM, the proposed method also showed higher performances by an accuracy of more than 2.2% for every set of conditions.

About the comparison of proposed phoneme HMMs with different candidate approaches, *i.e.*, the HMMs network of phoneme in the proposed model are concatenated using those phonemes with 1-best hypothesis and that with the correct transcription. We found that as the number of estimated sound source positions increased, the proposed method in use of the correct transcription became obviously better than that method with 1-best hypothesis. In comparison, for each number of training data (words), there were no significant differences between the performances of two candidate approaches.

Figs. 8(a) and (b) illustrate respectively the 7-th of the mel cepstral coefficients and 10-th order of that, where the ATFs are calculated from Eq. (9). As shown in the figures, there is the highest Fisher's ratio (*i.e.*, ratios of the within-class variances to the between-class variances) for each word in the case three positions case, which means the distribution of the ATF for each position can be classified easily. In comparison, the ratio of mean acoustic transfer func-



(a) The experimental room **(b)** Loudspeaker orientations

Figure 9: The illustration of experimental room and the loudspeaker orientations ($-45°$, $0°$, $45°$).

tion values for each word in the seven-position case, is highest, which means that it becomes difficult to classify the distribution for each sound source, when the number of source positions is seven. Therefore, based on these results, we can find that the tasks of verifying sound source localizations are deeply dependent on the accurate estimation of the ATF contribute, which is also supported by the comparison performance results between the proposed method with the 1-best hypothesis approach and that with the correct transcription.

6.3. Experimental Results in a Real World Environment

The proposed method was also evaluated in a real world environment. Fig. 9 (a) shows the experimental room environment and the position of the loudspeaker. The size of the recording room was about 6,300 micrometers (mm) × 3,200 mm × 2,800 mm (width × depth × height). The reverberation time was about 350 microseconds, and the SNR was about 41.49 [dB]. The distance from each position to the microphone was 1,500 mm. The speech signal was recorded using two (directional-type) microphones in order to provide a comparison with conventional CSP [46, 47] analysis, but the signal recorded by only one microphone (+ one camera) was used for the proposed method. The experimental devices were set on the table and desk. Three loudspeaker positions, *i.e.*, $-45°$, $0°$ and $45°$ were set for training and testing, and one loudspeaker was used

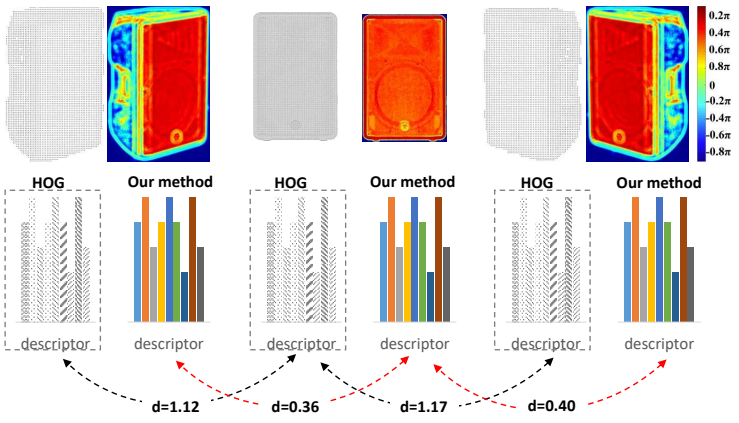


Figure 10: Comparison of our method and HOG: our method allows more flexible image representation, and produces smaller feature distances between the test and rotated object images.

for sound sources. Figure 9 (b) shows the differences in the orientation of the loudspeaker.

The speech sound data for experiments were from the ATR Japanese speech database. 2,620 words were used for training the clean speech sound HMM. We recorded 216 words as reverberant speech signal data for each sound source position, where 50 of them were used to train the ATF for each sound source position and the rest ones were test data. The accuracy calculation followed the rule of 4-fold cross validation. The test data contained 216 words $\times$ 3 persons who were randomly selected from the ATR database, totally had 648 words. The speech sound data used as the clean speech sound model training data, the ATF training data and testing data were spoken by the same person but with different speech contents. The other experimental setting are the same as those described in Subsection 6.1.

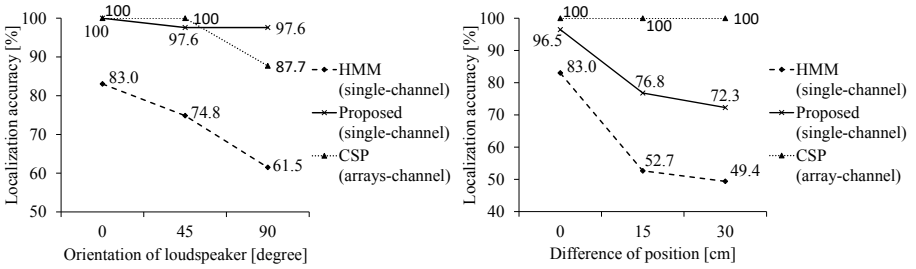
Few methods have been proposed for the task of the verification of sound localization information using the input of a single microphone. The method proposed in this paper was therefore compared with the CSP algorithm which

is a popular microphone-array-based method. Fig. 9(a) shows illustration of the experimental environment. The orientation for testing was changed to 0° (same to their statuses in the training stage), 45° and 90° (different from their statuses in the training stage). Similarly, the distance was changed to 0, 15 and 30 centimeters.

For visual model training (based on the same SVM framework), we respectively recorded 150 sample images of the sound source object (0 degree) at positions of 0, 15 and 30 centimeters. The sample images are normalized as the size of 100 × 200 pixels for training. Since the HOG-type feature is not scale-invariant, to enhance the generalization ability of the learning process, we performed a variety of scale transformations on the training samples, ultimately increasing the original number of samples by a factor of 10. Meanwhile, we also prepared images of the surrounding environment as examples of negative data. In the training stages, the positive-negative ratio is roughly 1:6.

Fig. 10 compares the distances in the feature space of the speaker for different orientations. In these cases, we aim at evaluating the feature similarity between the rotated image and the test image (0°). By this we can find the proposed image feature method produces smaller feature distances between the test and rotated images. This means the proposed Fourier domain HOG features can stably represent the image in different orientations. Consequently, our method has a relatively stable capacity to capture highly invariant features, *i.e.*, the representation of proposed feature is stable enough to resist the observed point shaking.

Fig. 11(a) shows accuracies of different orientations that were changed from the training stage by 0, 45, and 90 degrees. As shown in the Fig. 11(a), the comparison method based on CSP algorithm obtained the accuracy of 100% in the 0-degree and 45-degree cases. However, the accuracy of the case of 90 ° was 87.7%, because reflected signals (reverberation) from walls turned quite large. The similar situation also existed in the proposed phoneme HMM model (without visual model, marked as HMM (single-channel)). In comparison, the proposed multimodal fusion model (Proposed (signal-channel)) was slightly ef-



(a) The orientation-changed **(b)** The distance-changed

Figure 11: Accuracies of different orientation and distances changed from the training stage.

fected. These mean that the influence of reverberation signals obviously degrades the performance of both acoustic models using signal microphone and the microphone array. We also found that the proposed phoneme HMM is sensitive to the orientation parameter. Fig. 11(b) indicates accuracies of different distances that changed from the training stage by 0, 15, and 30 centimeters. As shown in Fig. 11(b), the accuracy of the acoustic model reduced dramatically when then changed distance is more than 15 centimeters, which means that the features of the ATF changed dramatically when the position was changed by 15 centimeters from the distance of trained stage.

From Figs. 11(a) and (b), we can find that the localization accuracy of the proposed acoustic model (marked as HMM) degraded as the orientation angle of the loudspeaker changed significantly. This means that the ATF depends not only on the position but also on the orientation of the speaker. Moreover, the features of the ATF changed from those used for training despite being measured from the same position. As shown in Fig. 11(a), the accuracy of the acoustic model degraded dramatically at the point where the difference between the positions of the loudspeaker for training and testing was 15 centimeters. Meanwhile, the CSP algorithm estimated the location with an accuracy of 100% for every condition. This means that the features of the ATF changed dramatically when the position was changed by 15 centimeters, although the phase difference used

in the CSP algorithm was stable and changed very little.

Based on these results, the conventional approaches that simultaneously extract and separate the parameters of localization (orientation and distance) suffer from the fact that these features influence each other. This limits their performance for the single-channel sound source localization task. By comparison, the new approach developed in this paper that individually estimates the distance and orientation of the sound source improves the level of achievable performance by about 22.7%.

7. CONCLUSION

This paper has used machine learning techniques to develop a novel solution to verify the sound source localization information using a single microphone. We have explored an effective location estimation framework based on fusing acoustic and visual models. The acoustic model uses the ATF to estimate the distance of sound scource. Specifically, the ATF is estimated using phoneme HMMs of clean speech sound together with a label sequence obtained from phoneme recognition. By so doing, the proposed method estimates the ATF more accurately than existing methods.

We also combine acoustic and visual models to help to identify speaker position. We have proposed a novel method based on Fourier domain analysis to efficiently implement the HOG-type polar feature descriptor. The proposed polar feature representation has rotation-invariant capabilities, which facilitates stable sound source object detection. We can therefore use the observed sound source object position to reliably calculate the orientation of the source object based using the convex imaging camera theory. The proposed combined visual-acoustic method improves the performance by about 22.7% for experiments conducted in real world acoustic environments. Therefore, the approach of multimodal fusion should be an important solution to the verification tasks of source localization information.

The proposed method is effective indoors and it has an impact on the rever-

berant environment. Future work will try to develop a deep model to capture more meticulous signal patterns to estimate the ideal ATF. Correspondingly, we also need to organize a larger dataset.

Acknowledgment

This work was supported in part by JSPS KAKENHI (Grant No. 17H01995 and 19H00597), Research Funds of State Grid Shaanxi Electric Power Company and State Grid Shaanxi Information and Telecommunication Company (contract no.SGSNXT00GCJS2000104).

References

[1] C. Zhang, D. Florencio, D. E. Ba, Z. Zhang, Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings, IEEE Trans. Multimedia (TMM) 10 (3) (2008) 538–548.

[2] K. Wu, D. Zhang, G. Lu, Z. Guo, Joint learning for voice based disease detection, Pattern Recognition 87 (2019) 130 – 139.

[3] K. Wu, V. G. Reju, A. W. H. Khong, S. T. Goh, Swarm intelligence based particle filter for alternating talker localization and tracking using microphone arrays, IEEE/ACM Trans. on Audio Speech and Language Processing (TASLP) 25 (6) (2017) 1384–1397.

[4] K. Wu, A. W. Khong, Sound source localization and tracking, in: Context Aware Human-Robot and Human-Agent Interaction, 2016, pp. 55–78.

[5] J. Li, H. Zhang, P. Wang, Blind separation of temporally correlated noncircular sources using complex matrix joint diagonalization, Pattern Recognition 87 (2019) 285 – 295.

[6] M. Baelde, C. Biernacki, R. Greff, Real-time monophonic and polyphonic audio classification from power spectra, Pattern Recognition 92 (2019) 82 – 92.

[7] Z. Wang, D. Wang, A joint training framework for robust automatic speech recognition, Vol. 24, 2016, pp. 796–806.

[8] X. Alameda-Pineda, R. Horaud, A geometric approach to sound source localization from time-delay estimates, IEEE/ACM Trans. on Audio Speech and Language Processing (TASLP) 22 (6) (2014) 1082–1095.

[9] M. Raspaud, H. Viste, G. Evangelista, Binaural source localization by joint estimation of ild and itd, IEEE Transactions on Audio, Speech, and Language Processing 18 (1) (2010) 68–77.

[10] H. Choi, S. Choi, Robust kernel isomap, Pattern Recognition 40 (3) (2007) 853 – 862.

[11] A. Fuchs, C. Feldbauer, M. Stark, Monaural sound localization, in: Proc. Interspeech 2011, Florence, Italy, 2011, pp. 2521–2524.

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

[13] R. Parhizkar, I. Dokmani, M. Vetterli, Single-channel indoor microphone localization, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 1434–1438.

[14] P. Kumarakulasingam, A. Agah, Neural network-based single sensor sound localization using a mobile robot, Intelligent Automation & Soft Computing 14 (1) (2008) 89–103.

[15] T. Takiguchi, Y. Sumida, Y. Ariki, Estimation of room acoustic transfer function using speech model, in: 2007 IEEE/SP 14th Workshop on Statistical Signal Processing, IEEE, 2007, pp. 336–340.

[16] N. Dalal, B. Triggs, Histograms of Oriented Gradients for Human Detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2005, pp. 886–893.

[17] J. Oliveira, C. Sousa, M. T. Coimbra, Coupled hidden markov model for automatic ecg and pcg segmentation, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 1023–1027.

[18] T. Takiguchi, S. Nakamura, K. Shikano, HMM-separation-based speech recognition for a distant moving speaker, IEEE/ACM Trans. on Audio Speech and Language Processing (TASLP) 9 (2) (2001) 127–140.

[19] T. Takiguchi, Y. Sumida, R. Takashima, Y. Ariki, Single-channel talker localization based on discrimination of acoustic transfer functions, EURASIP Journal on Advances in Signal Processing 2009 (2009) 9 pages.

[20] C. Rascon, G. Fuentes, I. Meza, Lightweight multi-doa tracking of mobile speech sources, EURASIP Journal on Audio, Speech, and Music Processing 2015 (1) (2015) 11.

[21] S. Ogiso, T. Kawagishi, K. Mizutani, N. Wakatsuki, K. Zempo, Self-localization method for mobile robot using acoustic beacons, ROBOMECH Journal 2 (1) (2015) 12.

[22] M. S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking, IEEE Transactions on Signal Processing 50 (2) (2002) 174–188.

[23] M. S. Brandstein, H. F. Silverman, A robust method for speech signal time-delay estimation in reverberant rooms, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1, IEEE, 1997, pp. 375–378.

[24] C. Knapp, G. Carter, The generalized correlation method for estimation of time delay, IEEE transactions on acoustics, speech, and signal processing 24 (4) (1976) 320–327.

[25] F. Deng, S. Guan, X. Yue, X. Gu, J. Chen, J. Lv, J. Li, Energy-based sound source localization with low power consumption in wireless sensor networks, IEEE Transactions on Industrial Electronics 64 (6) (2017) 4894–4902.

[26] S. Tian, U. Bhattacharya, S. Lu, B. Su, Q. Wang, X. Wei, Y. Lu, C. L. Tan, Multilingual scene character recognition with co-occurrence of histogram of oriented gradients, Pattern Recognition 51 (2016) 125 – 134.

[27] G. Takacs, V. Chandrasekhar, S. Tsai, D. Chen, R. Grzeszczuk, B. Girod, Fast computation of rotation-invariant image features by an approximate radial gradient transform, IEEE Trans. Image Proc.(TIP) 22 (8) (2013) 2970–2982.

[28] J. Chen, Z. Luo, Z. Zhang, F. Huang, Z. Ye, T. Takiguchi, E. R. Hancock, Polar transformation on image features for orientation-invariant representations, IEEE Trans. Multimedia (TMM) 21 (2019) 300 – 313.

[29] P. Smaragdis, P. Boufounos, Position and trajectory learning for microphone arrays, IEEE/ACM Trans. on Audio Speech and Language Processing (TASLP) 15 (1) (2007) 358–368.

[30] Y.-C. Lu, M. Cooke, Binaural distance perception based on direct-to-reverberant energy ratio, in: Proc. Int. Workshop Acoust. Echo Noise Control, 2008.

[31] Y.-C. Lu, M. Cooke, H. Christensen, Active binaural distance estimation for dynamic sources, in: Eighth Annual Conference of the International Speech Communication Association, 2007.

[32] E. Georganti, T. May, S. van de Par, A. Harma, J. Mourjopoulos, Speaker distance detection using a single microphone, IEEE transactions on audio, speech, and language processing 19 (7) (2011) 1949–1961.

[33] S. Vesa, Binaural sound source distance learning in rooms, IEEE Transactions on Audio, Speech, and Language Processing 17 (8) (2009) 1498–1507.

[34] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, Liblinear: A library for large linear classification, Journal of machine learning research 9 (Aug) (2008) 1871–1874.

[35] K. Liu, H. Skibbe, T. Schmidt, T. Blein, K. Palme, T. Brox, O. Ronneberger, Rotation-invariant hog descriptors using fourier analysis in polar and spherical coordinates, Int. J. Comput. Vis. (IJCV) 106 (3) (2014) 342–364.

[36] Y. Zhou, Q. Ye, Q. Qiu, J. Jiao, Oriented response networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), IEEE, 2017, pp. 4961–4970.

[37] M. Faundez-Zanuy, M. Hagmller, G. Kubin, Speaker identification security improvement by means of speech watermarking, Pattern Recognition 40 (11) (2007) 3027 – 3034.

[38] A. Sehr, R. Maas, W. Kellermann, Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition, IEEE/ACM Trans. on Audio Speech and Language Processing (TASLP) 18 (7) (2010) 1676–1691.

[39] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, B.-H. Juang, Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation, in: IEEE Int. Conf. Acoustics Speech and Signal Proc. (ICASSP), 2008, pp. 85–88.

[40] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, W. Kellermann, Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition, IEEE Signal Processing Magazine 29 (6) (2012) 114–126.

[41] R. Takashima, T. Takiguchi, Y. Ariki, Dimensional feature weighting utilizing multiple kernel learning for single-channel talker location discrimination

using the acoustic transfer function, The Journal of the Acoustical Society of America 133 (2) (2013) 891–901.

[42] B.-H. Juang, Maximum-likelihood estimation of mixture multivariate stochastic observations of Markov chains, AT&T Technical Journal 64 (6) (1985) 1235–1249.

[43] 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, 2001, pp. 43–46.

[44] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, K. Shikano, Atr japanese speech database as a tool of speech recognition and synthesis, Speech communication 9 (4) (1990) 357–363.

[45] T. Kinnunen, R. Saeidi, F. Sedlak, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, H. Li, Low-variance multitaper mfcc features: A case study in robust speaker verification, IEEE/ACM Trans. on Audio Speech and Language Processing (TASLP) 20 (7) (2012) 1990–2001.

[46] K. Kueviakoe, Z. Wang, A. Lambert, E. Frenoux, P. Tarroux, Localization of a vehicle: A dynamic interval constraint satisfaction problem-based approach, Journal of Sensors 2018.

[47] P. Jonsson, V. Lagerkvist, An initial study of time complexity in infinite-domain constraint satisfaction, Artificial Intelligence 245 (2017) 115–133.

Jinhui Chen received his Ph.D. degree (2016) in information science from Kobe University (Japan). From 2016 to 2020, he was an assistant professor at Kobe University. He is currently an associate professor at (Prefectural) University of Hiroshima (Japan). His research interests include pattern recognition (voice and image) and machine learning. He is a member of IEEE, ACM, and IEICE. He has published more than 20 publications in major journals and international conferences, such as IEEE Trans. Multimedia, IEEE/ACM Trans. Audio Speech Lang. Process., ACM MM, Interspeech etc.

Ryoichi Takashima received his degrees of B.S., M.Eng. and Ph.D. in information science at Kobe University in 2008, 2010, and 2013, respectively. He is currently an associate professor at Kobe University. His research interests include statistic signal processing and machine learning. He is a member of IEEE, IPSJ and ASJ.

Xingchen Guo received his B. Sc. degree (2015) in Mechanical Design Manufacture and Automation from Shaanxi University of Science & Technology, China. He is currently a Master student in the School of Management at Xi'an Jiaotong University, Xian, China. His research interests include pattern recognition and machine learning, particularly problems involving graphs and networks.

Zhihong Zhang received his BSc degree (1st class Hons.) in computer science from the University of Ulster, UK, in 2009 and the PhD degree in computer science from the University of York, UK, in 2013. He won the K. M. Stott prize for best thesis from the University of York in 2013. He is now an associate professor at the informatics school of Xiamen University, China. His research interests are wide-reaching but mainly involve the areas of pattern recognition and machine learning, particularly problems involving graphs and networks.

Xuexin Xu is currently a postgraduate student in the School of Informatics, Xiamen University. His research interests are computer vision and image synthesis.

Tetsuya Takiguchi received his B.S. degree in applied mathematics from Okayama University of Science, Okayama, Japan, in 1994, and his M.E. and Dr. Eng. degrees in information science from Nara Institute of

Science and Technology, Nara, Japan, in 1996 and 1999, respectively. From 1999 to 2004, he was a researcher at IBM Research, Tokyo Research Laboratory, Kanagawa, Japan. He is currently a professor at Kobe University. His research interests include statistic signal processing and pattern recognition. He received the award from the Acoustical Society of Japan in 2002. He is a member of IEEE, IPSJ and ASJ.

Edwin R. Hancock holds a BSc degree in physics (1977), a PhD degree in high-energy physics (1981) and a D.Sc. degree (2008) from the University of Durham, and a doctorate Honoris Causa from the University of Alicante in 2015. He is Professor in the Department of Computer Science, where he leads a group of some 25 faculty, research staff, and PhD students working in the areas of computer vision and pattern recognition. His main research interests are in the use of optimization and probabilistic methods for high and intermediate level vision. He is a fellow of the International Association for Pattern Recognition and the IEEE. He is currently Editor-in-Chief of the journal Pattern Recognition, and was founding Editor-in-Chief of IET Computer Vision from 2006 until 2012. He has also been a member of the editorial boards of the journals IEEE Transactions on Pattern Analysis and Machine Intelligence, Pattern Recognition, Computer Vision and Image Understanding, Image and Vision Computing, and the International Journal of Complex Networks.