Article:
guo, xingchen, Xu, Xuexin, Chen, Xunquan et al. (2022) Direction of Arrival Estimation for Indoor Environments Based on Acoustic Composition Model with a Single Microphone. Pattern Recognition. 108715. ISSN: 0031-3203

https://doi.org/10.1016/j.patcog.2022.108715

Direction of Arrival Estimation for Indoor Environments Based on Acoustic Composition Model with a Single Microphone

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

[a] *Xi'an University of Technology, Xi'an, China*
[b] *Xiamen University, Xiamen, China*
[c] *Kobe University, Kobe, Japan*
[d] *Prefectural University of Hiroshima, Hiroshima, Japan*
[e] *University of York, York, UK*

Abstract

This paper presents an effective method for multiple talker localisation using only a single microphone in a room. One of the main challenge here is obtaining a model that can be used for estimating the localization parameter. This model must be sensitive to all possible speaker locations and correctly discriminate their positions. The reverberant speech signal in a room environment can be composited by the clean speech and the acoustic transfer function (ATF). The ATF is a useful tool to describe changes of the speech source, and the approaches based on ATF can thus be used to identify talker localizations with a single microphone. This paper presents two methods, referred to as Composite Reverberant Speech (CRS) model and Direct Training Reverberant Speech (DTRS) model, and uses these methods for obtaining the ATF of a room. The approaches based on proposed methods can successfully and accurately process multi-talker localization task with single microphone. Experiments also demonstrate the effectiveness of the proposed methods.

Keywords: Gaussian Mixture Model (GMM), acoustic transfer function (ATF), talker localization

*Corresponding author: Xunquan Chen
Email address: cxq0720@hotmail.com (Xunquan Chen)

1. Introduction

The goal of talker localization is to automatically estimate the position of dominant talker which may alternate frequently among multiple candidate positions in the room environment. Effective methods for successful talker localization must be capable of simultaneously estimating the positions of all talkers present. This is an important capability in various applications in a number of pattern recognition areas, including disease detection [1, 2], human-machine interaction [3, 4], auditory scene analysis [5], augmented reality audio [6] and intelligent hearing aid devices [7]. Most of these applications require real-time processing of the signals. Further, the estimation of sound source location is frequently used in the subsequent processing stages, such as sound source separation [8, 9], sound source classification [10] and automatic speech recognition [11].

A common approach to talker localization relies on sound signal and utilizes time delay cues for localization by microphone arrays [3, 12]. It splits the microphone array on pairs and estimates the time difference of arrival (TDOA) at different microphone pairs, using *e.g.*, the generalized cross-correlation (GCC) algorithm [13]. The position of sound source can be determined by the triangulation rule given a set of signal frames via using some filter tools (*e.g.*, Bayesian filters) from different microphones pairs [14]. The TDOA-based method becomes unreliable when the individual TDOA estimates are inaccurate to begin with. A series of methods for improving the TDOA estimation problem have been proposed, where both the multipath and so far unexploited information among multiple microphone pairs were taken into account [15, 16, 17]. However, the microphone-array based systems are depend on large-size equipment and often computationally expensive, making them almost impracticable for real-time speech processing applications.

In this context, the ability to localize sound using a single microphone has emerged as an interesting, low-complexity and enlightening sound processing

domain. Indeed, the technique of single-channel sound source localization could be applied to a wider range of devices, especially for some small and low-power device with limited computational resources. For example, the wearable device and smart phone *etc.*, which would be quite important to some commerce applications and disaster relief tasks (*e.g.*, we can use mobile sensors to localize people who were buried under rubble in the earthquake by following their voice). Current examples of single microphone sound source location generally use the accessory information obtained from the external ear, such as head-related transfer functions (HRTFs), to localize the sound source [18, 19]. However, since they have to simultaneously extract and separate parameters of orientation and distance features, their accuracies are therefore still quite low. Other existing single microphone sound source location techniques mainly use learning-based mapping procedures, accompanied with the use of external pinnae and/or innerearcanals. For example, the work presented in [20] can only locate the types of sources for which it is trained for, therefore its performance might be affected by unknown sources. Further, a novel algorithm based on single microphone combined with an additional visual model is proposed recently [21].

Most of the work to date on the localization of sound sources is approached for the case of a single speech source, which is might not be appropriate in a number of situations. As in the cocktail-party scenario, for applications such as multi-conferencing, various gaming setups, and also human-computer interaction (HCI), it is often desirable to be able to distinguish between different simulataneous speakers. This requires sound source location to be extended to more complex and demanding problems. Recent work has focused on simultaneous speaker location [22, 23]. However, these methods were based solely on acoustic signal processing techniques and some utilized large-aperture microphone arrays. Therefore, the task of talker localization with a single microphone is still a challenging problem due to 1) the multiple voices of different talkers present, where an estimation model is required to switch from one talker to another, frequently using only a single-channel input; 2) the representation of positional characteristics obtained from signal frames which contain potential interference,

noise and reverberation.

To address these challenges, we propose novel approaches in this study. Generally, a reverberant speech can be linearly represented by the clean speech and acoustic transfer function (**ATF**) [24]. The **ATF** indicates transfer information of the speech signal in a room environment independent of the number of microphones. Therefore, even with single microphone, we also can capture the discriminative signal from different locations of speakers by using features based on the **ATF**. However, the **ATF** cannot be obtained directly, which has to be estimated from reverberant speech (also viewed as observed speech signal) using a clean speech model. In our study, we use Gaussian Mixture Model (GMM) to model clean speech features. By so doing, the approach can be independent of talker's utterance texts, because as widely known that it is difficult to obtain utterance texts in talker localization tasks. To estimate the **ATF** from reverberant speech, we proposed two approaches, *i.e.*, Composite Reverberant Speech (**CRS**) model and Direct Training Reverberant Speech (**DTRS**) model. In the **CRS** model, GMMs of the reverberant speech cepstrums are obtained by separately training clean speech GMM and single Gaussian model (SGM) (see Fig. 3) of the acoustic transfer function at each location in cepstrum domain, and then composing them (see Fig. 4). In **DTRS** model, the reverberant speech GMMs are directly trained from the speech signals recorded at each talker location in cepstrum domain (see Fig. 5). Finally, the **ATF** can be estimated from the obtained reverberant speech GMM by using trained clean speech GMM.

Based on above introduction, the main idea of our solution is to talker localization task that focuses on capturing the discriminative feature during the signal transferring from each location. In this way, this issue is reducible to a training task to get the **ATF**, which can be used to estimate the talker location. Meanwhile, it is not required to implement these approaches relying on a large number of signal filter or sensor tools. Since **ATF** is inherent information of speech and independent of the number of microphone, the approach based on the **ATF** can successfully process single microphone-based tasks. However, once the environmental conditions (*e.g.*, layout, wall material) change significantly,

ATF will become inappropriate for unknown indoor environment. Therefore, the main contribution of this paper is twofold, namely, a) this paper provides methods of **CRS** and **DTRS** and uses these to implement the **ATF** estimation; b) this paper proposes a novel solution using only a single microphone, which can accurately discriminate multi-talker locations in the room environment.

This paper is organized as follows: Section 2 reviews the related works. Sections 3 presents the proposed methods, including descriptions of **CRS** model, **DTRS** model, **ATF** estimation based on these models and technical details of talker localization implementation. These are followed by an experimental evaluation in Section 4. Finally, conclusions are presented in Section 5.

2. Related Works

Broadly, the existing localization methods can be divided into four main categories: a) time difference of arrival (TDOA) estimation based methods [16, 17, 25, 26, 27]; b) subspace-based methods adapted from classical spectrum estimation theory, such as multiple signal classification (MUSIC) [28, 29, 30, 31] and estimation of signal parameters via rotational invariance (ESPRIT) algorithms [32]; c) steered-response power (SRP) based methods [33, 34, 35]; d) independent component analysis (ICA) based methods [36, 37, 38, 39, 40].

Time difference of arrival (TDOA) is widely used for single source localization [41]. For systems with more than one microphone, we can first estimate the time TDOA among the signals captured by different microphones, using *e.g.*, the generalized cross-correlation (GCC) algorithm [13], steered-response power (SRP) [42] and its phase transferred version (SRP-PHAT) [43]. Then the position of sound source (speaker) can be determined by the triangulation rule given a set of TDOA's from different microphones pairs [14]. This basic bearing estimation process forms the foundation of most of the microphone-array based source-localization techniques, even though many algorithms may formulate and solve the problem from a different theoretical perspective [44]. Recently, several TDOA-based methods have also been proposed for multiple

source localization [26, 27]. However, the microphone-array based systems are depend on large-size equipment and often computationally expensive, making them almost impracticable for real-time speech processing applications. An overview of TDOA estimation techniques can be found in [41].

For multiple-speaker localization, some methods are able to localize a number of sound sources in overdetermined conditions (*i.e.*, sound sources number is equal or less than microphones number) [29], such as multiple signal classification (MUSIC) algorithm [28, 29, 30, 31]. MUSIC can estimate the directional of arrivals (DOAs) based on the eigen decomposition for the covariance matrix of observation vectors, but the accuracy is vulnerable to noise. Hu *et al.* [31] have proposed a novel MUSIC method by applying the relative sound pressure measurements of the higher-order microphone array in noisy environment. Another popular subspace-based method for DOA estimates is estimation of signal parameters by rotational invariance techniques (ESPRIT), which is more robust to array imperfections than MUSIC because it exploits the rotational invariance property in the signal subspace created by two subarrays [45]. However, these subspace methods in general, require a prior information on the number of active sources, which are often unavailable or difficult to obtain. Also while ESPRIT has a lower computational cost in comparison with MUSIC, it may still fail for directions where the estimation function is singular [46].

Maximizing the steered response power (SRP) of a beamformer is also used to estimate DOAs of multiple sources [33]. The main idea of the SRP is to steer the microphone array to all possible candidate source locations and find one where the response power is highest, typically using some frequency weighting. Furthermore, SRP with phase transform (SRP-PHAT) algorithm is also used for sound source localization [34, 35], which features robustness in noisy and reverberant environments. Although SRP based methods can provide excellent DOA estimation accuracy, two important problems prevent their widespread use in DOA estimation: a) computational cost due to performing a time-consuming search process over some space [47, 33], and b) robustness to additive noise [45].

Derived as a solution to the blind source separation (BSS), independent

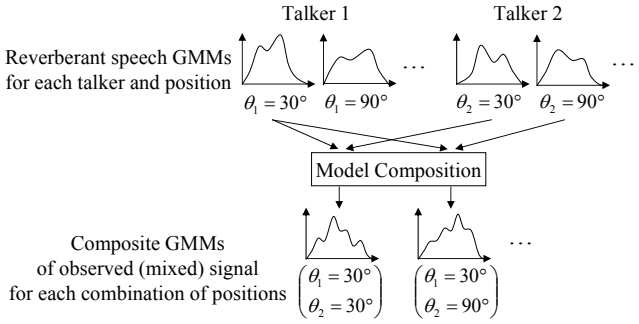

Figure 1: Training of mixed speech models of talkers using a model composition.

component analysis (ICA) methods achieve multiple source localization by directional sparsity of sound sources [36, 37, 38, 39]. The work of [36] proposes implementing ICA in regions of time-frequency domain for multiple sound sources localization and assumes that the number of sound sources is less than the number of microphones in each time-frequency region. Based on W-disjoint orthogonality (W-DO) assumption that only one source is active at each time-frequency domain, sparse component analysis (SCA) is applied for multiple sound source localization [48, 49]. Using this assumption, the problem of multiple sources localization might be solved by single source one for each time-frequency domain. For example, Pavlidi *et al.* [49] and Loesch *et al.* [50] presented an SCA-based method to count and localize multiple sound sources but requires one sound source to be dominant over others in a time-frequency zone. Most of the SCA-based methods are dependent on the W-DO property of multiple sound sources meaning that respective time-frequency representations of sources are located in different time-frequency zones. However, when the number of simultaneously occurring sources are four or above, more than one source is active in a time-frequency zone with a high probability. It means that the W-DO assumption of speech signal on which these methods are relying is less accurate with the number of actual sound sources raises, which would also affect the localization accuracy of the SCA-based method.

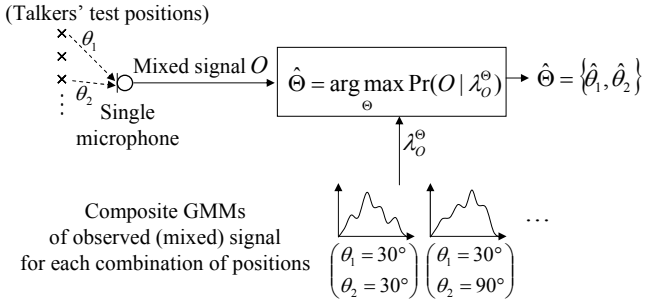

Figure 2: Talkers localization using composite models of the mixed speech.

3. the Estimation of Acoustic Transfer Function

In the study presented in this paper, the acoustic transfer function was estimated from observed (reverberant) speech using a clean speech model (speaker-dependent model). A Gaussian mixture model (GMM) was used to model the features of the clean speech. Because the characteristics of the acoustic transfer function depends on the talker's position, the obtained acoustic transfer function can be used to localize the talker. To estimate the location of sound sources without an external ear, we must extract the characteristics of each position. This is trained using training utterances at different positions. By using GMM source separation, we can estimate the acoustic transfer function with some adaptation data (only several words) uttered from different positions. Therefore, we can use the acoustic transfer function based on GMM to estimate locations of multiple talkers.

3.1. Overview of proposed method

An overview of the proposed method for localizing speakers using a single microphone is shown in Figures 1 and 2. The proposed method is divided into two steps: model training and localization. Figure 1 shows the model training section, and Figure 2 shows the localization section. The model training section first trains GMMs of the reverberant speech cepstrums of each speaker for all speaker locations. Afterwards, composite GMMs of the cepstrums of the mixed

speech signals of the speakers for all combinations of locations are obtained by combining these models to compose the model.

The normal distribution is a simple distribution function with a single peak and cannot represent complex distribution shapes. Therefore, a distribution function with multiple peaks is considered using weighted sums of multiple normal distributions, which is Gaussian Mixture Model.

If $\Pr(m) = w_m$ ($\sum_m w_m = 1$) is the prior probability (mixture weights) that the normal distribution (mixture elements) in the GMM with M mixtures will be output, the output probability (likelihood) of the GMM for the input can be expressed as:

$$\Pr(\mathbf{x}|\lambda) = \sum_{m=1}^{M} \Pr(m) \Pr(\mathbf{x}|m, \lambda) \quad (1)$$

$$= \sum_{m=1}^{M} w_m N(\mathbf{x}; \mu_m, \mathbf{\Sigma}_m) \quad (2)$$

where λ denotes the set of GMM parameters $\lambda = \{w_m, \mu_m, \mathbf{\Sigma}_m | m = 1, \cdots, M\}$. We can use the maximum likelihood estimation method for GMM parameter estimation, based on the training approach. When we estimate the parameters of a multidimensional normal distribution by the maximum likelihood estimation method, the formula for each parameter is obtained by differentiating the formula by each parameter and setting it to 0. The parameters are determined in such a way that the sum of the likelihoods of the models for each piece of data is maximised.

On the other hand, a mixture factor m exists in the GMM as a hidden variable. That is, the data $\mathbf{x}$ generated from the GMM alone is incomplete as data, and the data $(\mathbf{x}, m)$ is complete only when it is observed from which mixture element m it was generated. However, when learning a GMM using training data $\mathbf{x}$, the training data cannot be solved as simply as a normal distribution because the mixture element m that generates it is unknown and incomplete. Therefore, each parameter is estimated using mixed speech models.

The localization section calculates the likelihood of the evaluation speech

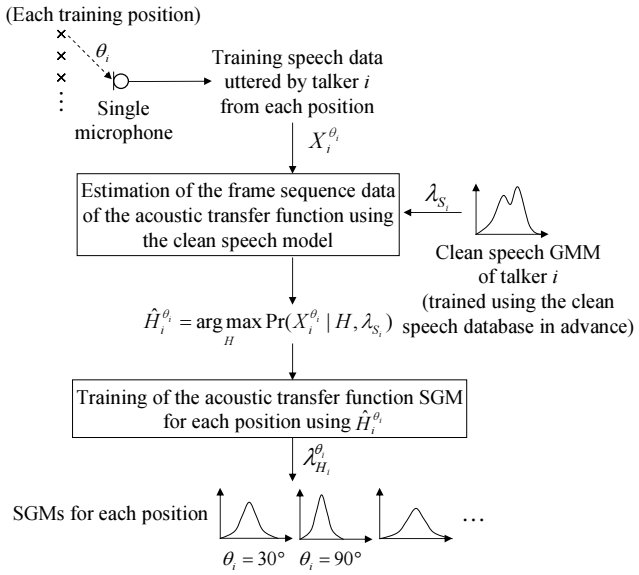

Figure 3: Training process for the acoustic transfer function SGM.

signal in which both speakers were speaking for the mixed speech signal models that were composed for each combination of locations. It then returns the combination of locations for which the likelihood was the highest as the locations of the speakers in the evaluation speech signal.

3.2. Composite reverberant speech model and direct training reverberant speech model

In this research, we propose two methods for obtaining GMMs of reverberant speech cepstrums for each speaker location in the model training stage as described in the previous subsection. In one method, which is referred to as the **CRS** model, GMMs of the reverberant speech cepstrums are obtained by separately training a GMM of the cepstrum of the clean speech signal and SGMs of the cepstrums of the **ATF** at each location, and then composing them. In the other method, which is referred to as the Directly-Trained Reverberant Model, the GMMs are trained directly from the speech signals for each speaker location. In both methods, the mixed speech GMMs are obtained by composing the

reverberant speech GMMs for each speaker. However, the **CRS** model differs from the **DTRS** model in that the **CRS** model creates the reverberant speech GMMs of each of the speakers by composing the clean speech GMM and the SGM of **ATF**.

The process for training the SGMs of the **ATF**s used in the **CRS** model is shown in Figure 3. First, in order to obtain the **ATF** that will be used for training, we record speech from a speaker at a designated location. Next, we estimate the cepstrums of the **ATF**s $\hat{H}_i^{\theta_i}$ from the recorded reverberant speech data $X_i^{\theta_i}$ $(i = 1, 2)$ using the GMMs of the clean speech cepstrums of the same speaker (λ_{S_i} represents the model parameters) through maximum likelihood estimation. Next, we learn the model parameters $\lambda_{H_i}^{\theta_i}$ of the SGM of the **ATF** at that particular location using the estimated **ATF** cepstrum. We perform this procedure for all sound source locations. In addition, we also train the **ATF** SGM for each position for the other speaker in the same way. Since it is normally assumed that the **ATF** does not depend on the speaker, it is not necessary to carry out this procedure separately for each speaker. However, in this research, we consider the possibility that the **ATF** may not necessarily be completely independent of the speaker. Therefore, we chose to train SGMs for the **ATF**s for each speaker.

The process for creating mixed speech GMMs for speakers, using model composition of the **CRS** models, is shown in Figure 4. All acoustic models are represented in the cepstral domain. In the method that uses **CRS** models, we first compose the obtained the SGMs of **ATF** and the clean speech GMM, which was used for estimating the **ATF**, in the cepstral domain in order to create the reverberant speech GMM (**CRS** model) for the given speaker at the given location. Next, applying the additivity of mixed speech in the linear spectral domain, we apply the inverse discrete cosine transform (IDCT) to the parameters $\lambda_{X_i}^{\theta_i}$ of the obtained reverberant speech GMMs of each speaker and transform them to the linear spectral domain. In the linear spectral domain, we compose the reverberant speech models of the speakers at the given locations, apply the log transform and the discrete cosine transform (DCT), and create

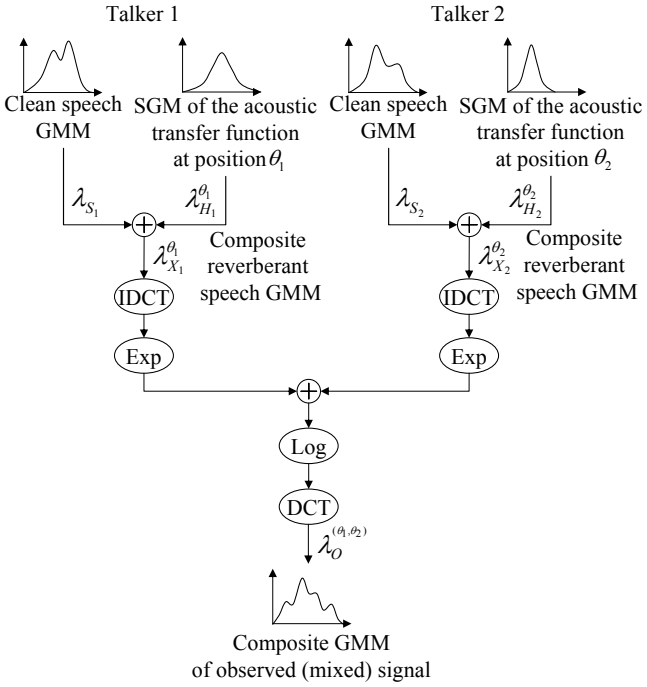

Figure 4: Composite model of the mixed speech of talkers using **CRS** model.

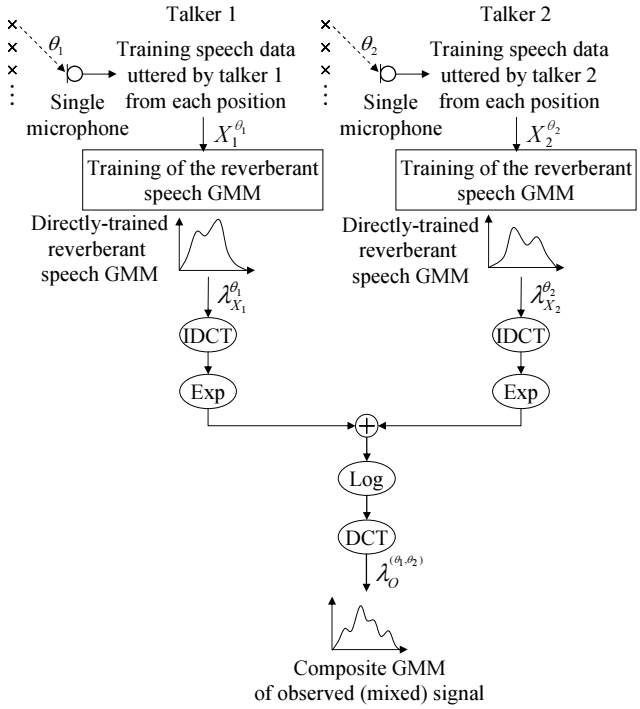

Figure 5: Composite model of the mixed speech of talkers using **DTRS** model.

the GMM (parameter $\lambda_O^{(\theta_1,\theta_2)}$) of the mixed speech cepstrum.

On the other hand, the procedure for creating mixed speech GMMs for the speakers through model composition of **DTRS** models is shown in Figure 5. In the method that directly uses **DTRS** models, we directly train GMMs (**DTRS** models) of the reverberant speech cepstrums based on the speech data from each location for each speaker. Afterwards, we compose the reverberant speech GMMs (**DTRS** models) of each speaker and create mixed speech GMMs, similar to the process we used in the method that uses **CRS** models.

We compose mixed speech GMMs for all combinations of locations. We calculate the likelihood of the evaluated mixed speech signal for the composed mixed speech models and return the pair of locations corresponding to the GMM for which the likelihood is the highest as the locations of the speakers. In the next section, we explain our method for estimating the **ATF**s used in the **CRS**

models.

3.3. Estimation of acoustic transfer functions

The reverberant speech data of each speaker is defined as $x(t)$. The cepstrum of the reverberant speech can be approximated by linear summation [24], as shown follows,

$$X_{\text{cep}}(d;n) \approx S_{\text{cep}}(d;n) + H_{cep}(d;n), \tag{3}$$

where X_{cep}, S_{cep}, and H_{cep} represent the cepstrums of the reverberant speech, clean speech, and **ATF**, and d represents the dimension, n denotes the frame index.

We estimated the **ATF** $\hat{H}_i^{\theta_i}(d;n)$ for when speaker i is speaking at location θ_i based on the reverberant speech data $X_i^{\theta_i}(d;n)$ and used it to train the SGM.

$$\mu^{(H_i^{\theta_i})} = \frac{1}{N}\sum_n H_i^{\theta_i}(n) \tag{4}$$

$$\Sigma^{(H_i^{\theta_i})} = \frac{1}{N}\sum_n (H_i^{\theta_i}(n) - \mu^{(H_i^{\theta_i})})^T (H_i^{\theta_i}(n) - \mu^{(H_i^{\theta_i})}) \tag{5}$$

$\mu^{(H_i^{\theta_i})}$ and $\Sigma^{(H_i^{\theta_i})}$ represent the mean vector and covariance matrix of the SGM of the **ATF**. Since the weak correlation among the different components of the cepstral vector, we assume that the covariance matrix is a diagonal matrix to reduce the training difficulty [51, 52].

The frame sequence of the acoustic transfer function in Equation 3 is estimated in an ML manner by using the expectation maximization (EM) algorithm, which maximizes the likelihood of the reverberant speech:

$$\hat{H}_i^{\theta_i} = \underset{H}{\operatorname{argmax}} \Pr(X_i^{\theta_i} | H, \lambda_{S_i}) \tag{6}$$

Here, λ_{S_i} denotes the set of clean speech GMM 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.

$$
\begin{aligned}
&Q(\hat{H}_i^{\theta_i} \mid H) \\
&\quad = E\left[\log \Pr\left(X_i^{\theta_i}, c \mid \hat{H}_i^{\theta_i}, \lambda_{S_i}\right) \mid H, \lambda_{S_i}\right] \\
&\quad = \sum_c \frac{\Pr\left(X_i^{\theta_i}, c \mid H, \lambda_{S_i}\right)}{\Pr\left(X_i^{\theta_i} \mid H, \lambda_{S_i}\right)} \cdot \log \Pr\left(X_i^{\theta_i}, c \mid \hat{H}_i^{\theta_i}, \lambda_{S_i}\right)
\end{aligned}
\tag{7}
$$

Here c represents the unobserved mixture component labels corresponding to the recorded reverberant speech $X_i^{\theta_i}$. The joint probability of observing sequences $X_i^{\theta_i}$ and c can be calculated as:

$$
\Pr\left(X_i^{\theta_i}, c \mid \hat{H}_i^{\theta_i}, \lambda_{S_i}\right) = \prod_{n^{(v)}} w_{n^{(v)}} \Pr\left(X_{i,n^{(v)}}^{\theta_i} \mid \hat{H}_i^{\theta_i}, \lambda_{S_i}\right), \tag{8}
$$

where w is the mixture weight and $X_{i,n^{(v)}}^{\theta_i}$ is the cepstrum at the n-th frame for the v-th training data. Since we consider the acoustic transfer function as additive noise in the cepstral domain, the mean to mixture k in the model $\lambda_{X_i^{\theta_i}}$ is derived by adding the acoustic transfer function. Therefore, Equation 8 can be written as:

$$
\begin{aligned}
&\Pr\left(X_i^{\theta_i}, c \mid \hat{H}_i^{\theta_i}, \lambda_{S_i}\right) \\
&\quad = \prod_{n^{(v)}} w_{c_{n^{(v)}}} \cdot N\left(X_{i,n^{(v)}}^{\theta_i}; \mu_{n^{(v)}}^{(S_i)} + \hat{H}_{i,n^{(v)}}^{\theta_i}, \Sigma_{k_{n^{(v)}}}^{(S_i)}\right),
\end{aligned}
\tag{9}
$$

where $N(X_i^{\theta_i}; \mu, \cdot)$ denotes the multivariate Gaussian distribution. It is straightforward to derive that [53] .

$$
\begin{aligned}
&Q(\hat{H}_i^{\theta_i} \mid H) \\
&\quad = \sum_k \sum_{n^{(v)}} \Pr\left(X_{i,n^{(v)}}^{\theta_i}, c_{n^{(v)}} = k \mid \lambda_{S_i}\right) \log w_k \\
&\quad + \sum_k \sum_{n^{(v)}} \Pr\left(X_{i,n^{(v)}}^{\theta_i}, c_{n^{(v)}} = k \mid \lambda_{S_i}\right) \\
&\quad \cdot \log N\left(X_{i,n^{(v)}}^{\theta_i}; \mu_k^{(S_i)} + \hat{H}_{i,n^{(v)}}^{\theta_i}, \Sigma_k^{(S_i)}\right)
\end{aligned}
\tag{10}
$$

Here $\mu_k^{(S_i)}$ and $\Sigma_k^{(S_i)}$ are the k-th mean vector and the (diagonal) covariance matrix in the clean speech GMM, respectively. It is possible to train those parameters by using a clean speech database. Next, we focus only on the term involving H.

$$
\begin{aligned}
&Q(\hat{H}_i^{\theta_i} \mid H) \\
&= \sum_k \sum_{n^{(v)}} \Pr\left(X_{i,n^{(v)}}^{\theta_i}, c_{n^{(v)}} = k \mid \lambda_{S_i}\right) \\
&\cdot \log N\left(X_{i,n^{(v)}}^{\theta_i}; \mu_k^{(S_i)} + \hat{H}_{i,n^{(v)}}^{\theta_i}, \Sigma_k^{(S_i)}\right) \\
&= -\sum_k \sum_{n^{(v)}} \gamma_{k,n(v)} \sum_{d=1}^{D} \left\{ \frac{1}{2} \log(2\pi)^D {\sigma_{k,d}^{(S_i)}}^2 \right. \\
&\left. + \frac{\left(X_{i,n^{(v)},d}^{\theta_i} - \mu_{k,d}^{(S_i)} - \hat{H}_{i,n^{(v)},d}^{\theta_i}\right)^2}{2{\sigma_{k,d}^{(S_i)}}^2} \right\}
\end{aligned}
\tag{11}
$$

$$
\gamma_{k,n^{(v)}} = \Pr\left(X_{i,n^{(v)}}^{\theta_i}, k \mid \lambda_{S_i}\right) \tag{12}
$$

Here D is the dimension of the reverberant speech vector $X_{i,n^{(v)},n}^{\theta_i}$, and $\mu_{k,d}^{(S_i)}$ and ${\sigma_{k,d}^{(S_i)}}^2$ are the d-th mean value and the d-th diagonal variance value of the k-th component in the clean speech GMM, respectively. The maximization step (M-step) in the EM algorithm becomes "max $Q(\hat{H}_i^{\theta_i} \mid H)$". The re-estimation formula can, therefore, be derived, knowing that $\partial Q(\hat{H}_i^{\theta_i} \mid H)/\partial \hat{H}_i^{\theta_i} = 0$ as

$$
\hat{H}_{i,n^{(v)},d}^{\theta_i} = \frac{\sum_k \gamma_{k,n^{(v)}} \frac{X_{i,n^{(v)},d}^{\theta_i} - \mu_{k,d}^{(S)}}{{\sigma_{k,d}^{(S)}}^2}}{\sum_k \frac{\gamma_{k,n^{(v)}}}{{\sigma_{k,d}^{(S)}}^2}} \tag{13}
$$

After calculating the frame sequence data of the acoustic transfer function for all training data, the GMM for the acoustic transfer function is created.

3.4. Creation of mixed speech models through model composition

In the **CRS** model method, an SGM of the **ATF** is trained for each speaker and each location using the **ATF** $\hat{H}_i^{\theta_i}$ calculated in the previous subsection.

Then, the parameters λ_O^{Θ} ($\Theta = \{\theta_1, \ldots, \theta_M\}$) of the mixed speech GMMs for multiple speakers are calculated through model composition using the parameters $\lambda_{H_i}^{\theta_i} = \left\{\mu^{(H_i^{\theta_i})}, \Sigma^{(H_i^{\theta_i})}\right\}$ of the obtained **ATF** cepstrum SGMs and the parameters λ_{S_i} of the clean speech cepstrum GMMs for each speaker. M represents the number of speakers. The parameters $\mu_k^{(S)}$, $\Sigma_k^{(S)}$, $\mu^{(H_i^{\theta_i})}$ and $\Sigma^{(H_i^{\theta_i})}$ of the clean speech GMM and the **ATF** GMMs, described in the previous subsection, are represented as $\mu_{\text{cep,k}}^{(S)}$, $\Sigma_{\text{cep,k}}^{(S)}$, $\mu_{\text{cep}}^{(H_i^{\theta_i})}$, and $\Sigma_{\text{cep}}^{(H_i^{\theta_i})}$ in this section in order to emphasize the fact that they refer to models in the cepstral domain, where the subscript k denotes k-th Gaussian component.

The mixed speech spectrum is expressed as the linear sum of the reverberant speech of each speaker as shown follows.

$$O_{\text{spc}}^{\Theta}(\omega; n) = \sum_{i=1}^{M} X_{\text{spc,i}}^{\theta_i}(\omega; n); \tag{14}$$

$$X_{spc,i}^{\theta_i}(\omega; n) \approx S_{\text{spc,i}}(\omega; n) \cdot H_{\text{spc,i}}^{\theta_i}(\omega; n), \tag{15}$$

$S_{\text{spc,i}}(\omega; n)$ refers to the clean speech spectrum of speaker i. $X_{\text{cep,i}}^{\theta_i}(\omega; n)$ and $H_{\text{spc,i}}^{\theta_i}(\omega; n)$ denote the reverberant speech spectrum and **ATF** spectrum of the speaker at location θ_i and $O_{\text{spc}}^{\Theta}(\omega; n)$ represents the mixed speech spectrum when each speaker speaks from their given location.

In the **CRS** model method, we first calculate the reverberant speech model parameters using the model parameters of the clean speech model and the **ATF** model which were modeled in the cepstral domain. Since the reverberant speech data can be represented as a linear summation of the clean speech data and the **ATF** in the cepstral domain according to Equation 3, the mean vector ($\mu_{\text{cep,k}}^{(X_i^{\theta_i})}$) and covariance matrix ($\Sigma_{\text{cep,k}}^{(X_i^{\theta_i})}$) of the reverberant speech can be represented as linear summations of the mean vector and covariance matrix of the clean speech and **ATF** [54].

$$\mu_{\text{cep,k}}^{(X_i^{\theta_i})} = \mu_{\text{cep,k}}^{(S_i)} + \mu_{cep,k}^{(H_i^{\theta_i})}; \tag{16}$$

$$\Sigma_{\text{cep,k}}^{(X_i^{\theta_i})} = \Sigma_{\text{cep,k}}^{(S_i)} + \Sigma_{\text{cep,k}}^{(H_i^{\theta_i})}. \tag{17}$$

In the case of the **DTRS** model, the model parameters of the reverberant speech model are obtained by directly training the GMMs using the reverberant speech data.

$$\mu_{\mathrm{cep,k}}^{(X_i^{\theta_i})} = \sum_n \frac{\gamma_k(n) X_i^{\theta_i}(n)}{\sum_n \gamma_k(n)} \tag{18}$$

$$\Sigma_{\mathrm{cep,k}}^{(X_i^{\theta_i})} = \sum_n \frac{\gamma_k(n)(X_i^{\theta_i}(n)-\mu_{\mathrm{cep,k}}^{(X_i^{\theta_i})})^T(X_i^{\theta_i}(n)-\mu_{\mathrm{cep,k}}^{(X_i^{\theta_i})})}{\sum_n \gamma_k(n)} \tag{19}$$

where $\gamma_k(n)$ denotes posterior distribution. Note that in this research, we also assume that the covariance matrix is diagonal for the case of the **DTRS** model as well.

Next, we calculate the model parameters λ_O^{Θ} of the mixed speech signal using the model parameters of the reverberant speech model for each speaker obtained from Equation 16 and 17 or from Equation 18 and 19. Since the mixed speech signal can be represented as the linear summation of the reverberant speech signals of each speaker in the spectral domain according to Equation 14, we first transform $\lambda_{X_i}^{\theta_i}$ from the cepstral domain to the linear spectral domain. The transformation of the model parameters from the cepstral domain to the logarithmic spectral domain can be calculated by applying the inverse discrete cosine transform to each normal distribution in the GMM.

$$\mu_{\mathrm{log}}^{(X_i^{\theta_i})} = \Gamma^{-1}\mu_{\mathrm{cep}}^{(X_i^{\theta_i})} \tag{20}$$

$$\Sigma_{\mathrm{log}}^{(X_i^{\theta_i})} = \Gamma^{-1}\Sigma_{\mathrm{cep}}^{(X_i^{\theta_i})}(\Gamma^{-1})^T \tag{21}$$

where Γ represents the transformation matrix for the discrete cosine transform. $\mu_{\mathrm{log}}^{(X_i^{\theta_i})}$ and $\Sigma_{\mathrm{log}}^{(X_i^{\theta_i})}$ represent the mean vectors and covariance matrices of each reverberant speech signal in the logarithmic spectral domain. Here, the covariance matrix of each of the models is defined as a diagonal matrix in the cepstral domain. However, after undergoing the inverse discrete cosine transform, the covariance matrix is no longer a diagonal matrix in the logarithmic spectral

domain and has nonzero values for the covariance in the non-diagonal components. The computations in the logarithmic spectral domain and the linear spectral domain in the following sections are performed while taking this into consideration.

Next, the model parameters in the logarithmic spectral domain are transformed to the linear spectral domain. In this research, we assume that the model follows a normal distribution in the cepstral domain and in the logarithmic spectral domain which is a linear transform of the cepstral domain. If it is assumed that the logarithmic spectrum follows a normal distribution, then it is assumed that the linear spectrum, which is an exponential transform of the logarithmic spectrum, follows a log-normal distribution. A log-normal distribution is a distribution in which the logarithm of the variable follows a normal distribution. If we assume that the linear spectrum follows the log-normal distribution, then the mean and covariance matrix can be obtained as follows using the mean and covariance matrix of the normal distribution in the logarithmic spectral domain.

$$\mu_{\mathrm{spc,p}}^{(X_i^{\theta_i})} = \exp\left\{\mu_{\mathrm{log},p}^{(X_i^{\theta_i})} + \sigma_{\mathrm{log},pp}^{(X_i^{\theta_i})^2}/2\right\}; \tag{22}$$

$$\sigma_{spc,pq}^{(X_i^{\theta_i})^2} = \mu_{\mathrm{spc,p}}^{(X_i^{\theta_i})} \cdot \mu_{\mathrm{spc,q}}^{(X_i^{\theta_i})} \cdot \left\{\exp\left(\sigma_{\mathrm{log},pq}^{(X_i^{\theta_i})^2}\right) - 1\right\}, \tag{23}$$

where $\mu_{\mathrm{spc,p}}^{(X_i^{\theta_i})}$ and $\sigma_{\mathrm{spc,pq}}^{(X_i^{\theta_i})^2}$ represent the p-th element of the mean vector and the (p,q)-th element of the covariance matrix in the linear spectral domain, respectively. Next, we compose the model parameters of the mixed speech signal in the linear spectral domain from the model parameters of the reverberant speech for each speaker. Here, the mean vector ($\mu_{\mathrm{spc,k}}^{(O^{\Theta})}$) and the covariance matrix ($\Sigma_{\mathrm{spc,k}}^{(O^{\Theta})}$) of the mixed speech signal can be approximated as the linear summation of the mean vectors and the covariance matrices of the reverberant speech signals of each speaker in the linear spectral domain [55].

$$\mu_{\mathrm{spc,k}}^{(O^{\Theta})} \approx \sum_{i=1}^{M} \mu_{\mathrm{spc,k}}^{(X_i^{\theta_i})}, \quad \Sigma_{\mathrm{spc,k}}^{(O^{\Theta})} \approx \sum_{i=1}^{M} \Sigma_{\mathrm{spc,k}}^{(X_i^{\theta_i})} \tag{24}$$

Afterwards, we use the parameters of the mixed speech signal model (assumed to be a log-normal distribution) to calculate the model parameters in the log-

arithmic spectral domain (assumed to be a normal distribution). This can be calculated by carrying out the inverse of the procedure in Equations 22 and 23.

$$\sigma_{\log,pq}^{(O^\Theta)^2} = \log\left\{\frac{\sigma_{spc,pq}^{(O^\Theta)^2}}{\mu_{\mathrm{spc,p}}^{(O^\Theta)}\cdot\mu_{\mathrm{spc,q}}^{(O^\Theta)}}+1\right\} \quad (25)$$

$$\mu_{\log,p}^{(O^\Theta)} = \log\mu_{\mathrm{spc,p}}^{(O^\Theta)} - \sigma_{\log,pp}^{(O^\Theta)^2}/2 \quad (26)$$

Finally, we apply the discrete cosine transform and transform the data to the cepstral domain.

$$\mu_{\mathrm{cep}}^{(O^\Theta)} = \Gamma\mu_{\log}^{(O^\Theta)}, \qquad \Sigma_{\mathrm{cep}}^{(O^\Theta)} = \Gamma\Sigma_{\log}^{(O^\Theta)}\Gamma^T \quad (27)$$

In this study, after transforming the data back to the cepstral domain, we set the non-diagonal elements of the covariance matrix to be zero and redefine the matrix as a diagonal matrix.

3.5. Speaker localization by maximum likelihood criterion

Using the method described in the previous subsections, we first calculate GMMs for the mixed speech signals for all combinations of speaker locations in advance. Then we derive the likelihood of the composed mixed speech models for the mixed speech signal under evaluation and return the pair of locations that correspond to the GMM with maximum likelihood as the locations of each of the speakers.

$$\hat{\Theta} = \mathrm{argmax}_\Theta \Pr(\mathrm{O}|\lambda_{\mathrm{O}}^\Theta), \quad (28)$$

λ_O^Θ refers to the composed model of the mixed speech signal in the cepstral domain for the combination of positions denoted by Θ.

4. Experiments

4.1. Experiment environment

In order to evaluate the proposed method, we performed a talker localization experiment involving a given speaker. For the speech data, we used data

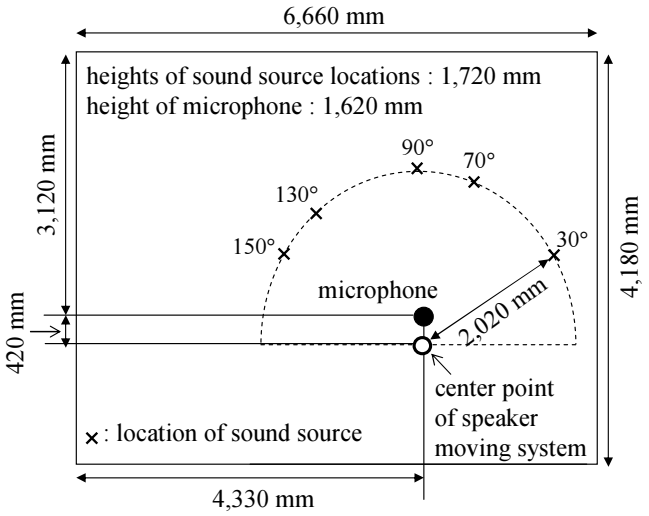

Figure 6: Experiment room environment for simulation.

of both male and female speakers randomly obtained from the ASJ research continuous speech database (ASJ-JIPDEC) [56], and the number of speakers is equal to a value of 2 in this research. We created reverberant speech signals for each speaker by convolving the speech data with the impulse response that was stored in the RWCP actual-environment speech and acoustics database [56]. The reverberation time was 300 milliseconds. We conducted the experiment for the case in which there were three sound source locations at 30°, 90°, and 150° and the case in which there were five sound source locations at 30°, 70°, 90°, 130°, and 150°. Under these conditions, there were 9 combinations and 25 combinations of locations of the speakers, respectively. Figure 6 shows the sound source locations used in the experiment described in this section.

The speech signal was analyzed using a sampling frequency of 12 kHz, a window width of 32 msec, and a frame shift of 8 msec. A 16-dimensional vector of MFCCs was used as the features. In the **CRS** model, the GMM for the clean speech signal of the designated speaker that was used for **ATF** estimation and reverberant speech model composition was trained using 40 sentences for each speaker. The number of mixture components in GMM was 64. The experiment

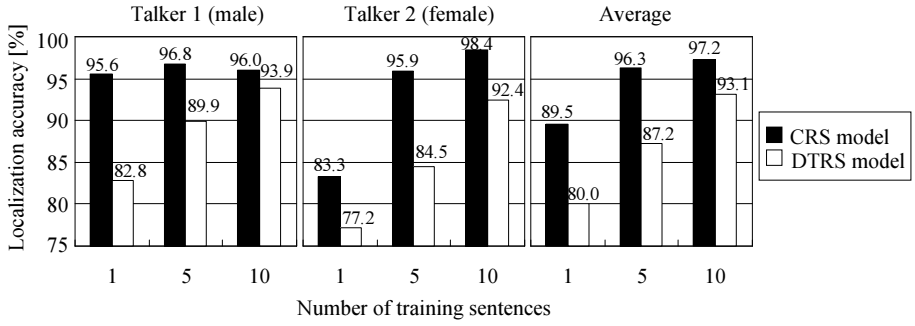

Figure 7: Single-talker localization accuracies (three positions) [%].

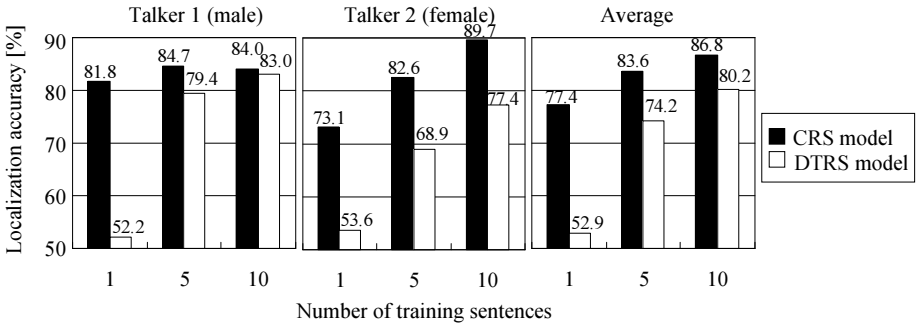

Figure 8: Single-talker localization accuracies (five positions)[%]

was performed using three different values for the number of sentences, which were 1 sentence, 5 sentences, and 10 sentences, used for training the **ATF** SGMs in the **CRS** models and the reverberant speech GMMs in the **DTRS** models. The evaluation was performed using 100 1-second segments of speech. The ratio of the power of the speakers in the evaluated mixed speech signal was an average of approximately ±5.90 dB per segment, with a standard deviation of 2.54 dB. We returned the combination of locations that correspond to the mixed speech GMM with the highest likelihood for each segment and evaluated the accuracy rate. Note that the content of the data used for learning the clean speech signal, the data used for learning the locations, and the data used for evaluation were different from each other.

4.2. Experiment results

The difference between the **CRS** model and the **DTRS** model is the fact that the **CRS** model calculates the model of the reverberant speech signal for a single speaker by composing the clean speech model and the **ATF**, while the **DTRS** model calculates the model by training it directly based on the reverberant speech. In order to evaluate the effect of this difference, we performed sound source localization experiments for a single speaker independently for each speaker. We calculated the likelihood of the evaluation data (reverberant speech) spoken by a single speaker for the reverberant speech models $\lambda_{X_i}^{\theta_i}$ for each location that were calculated using both methods and returned the location that had the highest likelihood. We evaluated the accuracy rate of this process. We assume that it is known which speaker is speaking.

The accuracy rate of the sound source localization for the case in which there were 3 locations and the case in which there were 5 locations are shown in Figure 7 and Figure 8. The figures show that the **CRS** model has a higher accuracy rate overall. In addition, the difference in the accuracy rate increases as the number of sentences used for training the reverberant speech models for each location decreases. This is because the **DTRS** model was unable to learn the different variations in the phonemes in the reverberant speech based on small amounts of training data. On the other hand, in the **CRS** model, the reverberant speech model was created by training the clean speech GMM beforehand using 40 sentences and composing them. Therefore, the **CRS** model was able to mitigate the amount of reduction in the accuracy even with small amounts of training data.

In the case in which there were 5 locations for the sound source, the number of classes for detection is larger, and the distance between locations decreases. Therefore, the localization accuracy rate is lower overall. However, in the case in which the number of sentences in the training data for the reverberant speech model was 1 sentence, the **DTRS** model had a 27.3 % decrease in the accuracy rate on average compared to the case in which 10 sentences were used, while the **CRS** model had only a 9.4% decrease. This result shows that the robustness

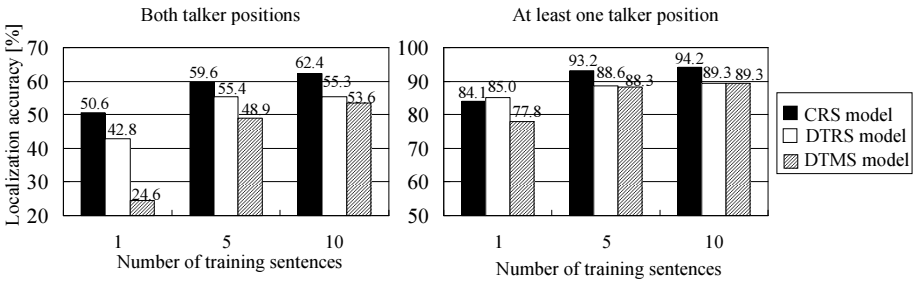

Figure 9: Two-talker localization accuracies [%], where the number of positions is three. Test data consists of 100 speech segments having a time length 1 sec.

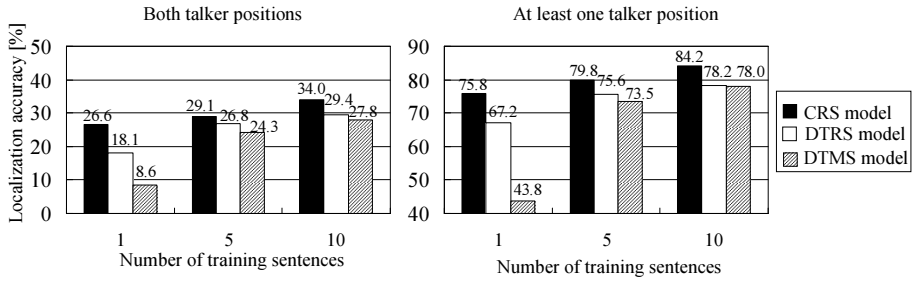

Figure 10: Two-talker localization accuracies [%], where the number of positions is five. Test data consists of 100 speech segments having a time length 1 sec.

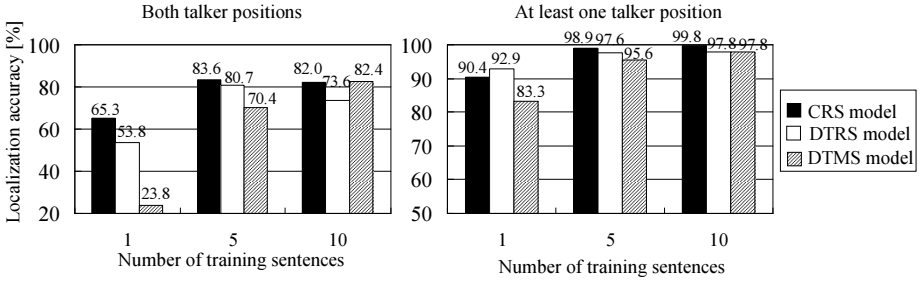

Figure 11: Two-talker localization accuracies [%], where the number of positions is three. Test data consists of 50 speech segments having a time length 5 sec.

against a decrease in the amount of training data is preserved even when the number of locations increases. Since this method uses speech spoken by the user in order to create the reverberant speech model, it is desirable for the amount of training data per location to decrease as the number of locations increases. While it is true that the **CRS** model requires several tens of sentences in order to train the clean speech GMM, this amount does not depend on the number of locations. Therefore, this method is more effective as the number of locations increases. On the other hand, in cases in which the number of locations is small, it is possible that the amount of overall training data required in order to directly learn the reverberant speech in the **DTRS** model could be smaller than the amount of training data required in the **CRS** model.

Table 1 shows the comparison of results among for the proposed and the existing methods. Since relatively few methods have been proposed for sound source localization using a single microphone, the method proposed was also compared with the CSP algorithm [57]. Experimental results show that the proposed methods based on both CRS and DTRS outperform existing method [58], and are even competitive with the microphone array-based approach (CSP).

Figures 9 and 10 show the localization accuracy results for the cases of 3 and 5 locations, respectively. In these experiments, we compared the **CRS** model, the **DTRS** model, and an additional method, in which the mixed speech GMM is trained without performing model composition. In the following sections, we

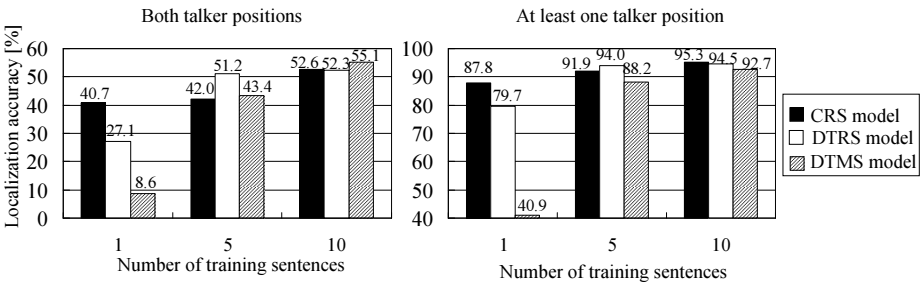

Figure 12: Two-talker localization accuracies [%], where the number of positions is five. Test data consists of 50 speech segments having a time length 5 sec.

Table 1: Comparison results with the proposed and existing methods (10 sentences) (%).

Method	Three positions	Five positions
Takiguchi *et al* [58]	84.1	64.1
CRS (Proposed)	**97.2**	**86.8**
DTRS (Proposed)	93.1	80.2
CSP (*microphone array) [57]	99.1	98.9

refer to this model as the *Directly-Trained Mixed Speech (DTMS*) model. In the DTMS model, we record the speech of speakers at different locations speaking simultaneously and train the GMMs directly using the resulting mixed speech. This method is capable of training mixed speech GMMs without performing model composition. In addition, "Both talker positions" represents the case in which both speakers were localized correctly, and "At least one talker position " denotes the case in which at least one speaker (as well as the case of both speakers) was localized correctly.

If we let W represent the number of utterances for training for each location and Y represent the number of locations, then the number of actual utterances per user is $W \times Y(+40)$ for the **CRS** model, $W \times Y$ for the **DTRS** model, and $W \times (Y^2)$ for the DTMS model, respectively. In addition, the expected value for the proportion of cases in which both speakers are localized correctly is 11.1% for the case of that there were 3 locations and 4.0% for the case in which there were 5 locations. The expected value for the proportion of cases is 55.6%, when at least one speaker is localized correctly (3 locations for test) and 36.0% for the case of 5 test locations.

Figures 9 and 10 show that the **CRS** model achieves a higher accuracy rate than the other models for almost all conditions. Since the model considers the combinations of phonemes in the mixed speech of speakers, the number of variations is larger than the case of speech from only one speaker. Therefore, the **CRS** model achieves a high accuracy rate because it uses a sufficient amount of data to train acoustic models for each speaker beforehand. A comparison between the **DTRS** model and the DTMS model shows the accuracy rate differs between the two models in cases in which the amount of training data is small. However, the proportion of cases in which both speakers were localized correctly was at most only 62.4% for the case where there were 3 locations, and 34.0% for the case in which there were 5 locations. This accuracy rate is not considered very high in terms of practical application.

Next, we changed the evaluation speech signal to consist of 50 segments of 5-second long speech signals and performed the experiment. The sound source

localization accuracy rate is shown in Figures 11 and 12. The figures show that the accuracy rate has increased compared to the results for 5-second long evaluation speech signals. The accuracy rates for the **DTRS** and DTMS models increased by a large amount compared to the **CRS** model for the cases in which the amount of training data was large in particular. This is because the increase in the length of the evaluation speech signals caused an increase in the amount of information, which made it possible for models that were able to represent mixed speech somewhat well, such as the **DTRS** model and DTMS model trained using 10 sentences, to localize speakers with a relatively high degree of accuracy.

5. Conclusions

In this paper, we have proposed solutions to the localization task of talkers. We address acoustic transfer function (**ATF**) which indicates changes of the speech signal in a room independent of the number of microphones, and it thus can be used to discriminate talker locations in single microphone tasks. In this study, we estimate the **ATF** from reverberant speech, using clean speech model. We process the speech signal in the cepstrum domain, and propose Composite Reverberant Speech (**CRS**) model and Direct Training Reverberant Speech (**DTRS**) model to obtain reverberant speech model. In our study, we use Gaussian Mixture Model (GMM) to model clean speech features, because the clean speech GMM is independent of talker's utterance texts, we therefore can process this task easily. Experiments are carried out to evaluate our methods, which shows the effectiveness of the proposed methods. Therefore, the ATF is a useful tool for estimating the localization task of talkers and we can obtain ATF of a room from the observed signal (reverberant speech) model using the clean speech model.

Although our new method offers some strong advantages, we will focus our future work on further improving it. First, the ATF estimation using Gaussian Mixture Model is not optimal, and this method has a great impact on the

quality of the reverberant environment. We still need a better feature model to more accurately represent the transfer function, and this will further improve the performance of estimation results. Second, there are theoretically superior methods for determining the location parameters, which should be data-independent. Third, our model needs to be retrained when the environmental conditions change significantly. Therefore, the generalization ability of the proposed model is indeed insufficient, and there will be some limitations in the scope of application. One of the most important pieces of future work will be to find the essential factors underlying effective speaker localization and develop improved representation models for them. Additionally, we will explore a more robust and general method for our future work.

Acknowledgment

This work was supported in part by JSPS KAKENHI (Grant No. 17H01995 and 19H00597), the National Natural Science Foundation of China under Grant 62176227, U2066213 and 61860206004. Xingchen Guo and Xuexin Xu are co-first authors, which contributed equally to this paper.

References

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

[2] T. K. Dash, S. Mishra, G. Panda, S. C. Satapathy, Detection of covid-19 from speech signal using bio-inspired based cepstral features, Pattern Recognition 117 (2021) 107999.

[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, Springer, 2016, pp. 55–78.

[5] G. J. Brown, M. Cooke, Computational auditory scene analysis, Computer Speech & Language 8 (4) (1994) 297–336.

[6] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, G. Lorho, Augmented reality audio for mobile and wearable appliances, Journal of the Audio Engineering Society 52 (6) (2004) 618–639.

[7] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, U. Rass, Signal processing in high-end hearing aids: state of the art, challenges, and future trends, EURASIP Journal on Applied Signal Processing 2005 (2005) 2915–2929.

[8] T. Jenrungrot, V. Jayaram, S. Seitz, I. Kemelmacher-Shlizerman, The cone of silence: Speech separation by localization, Advances in Neural Information Processing Systems 33 (2020) 20925–20938.

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

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

[11] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, D. Yu, Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition, Computer Speech & Language 75 (2022) 101360.

[12] 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.

[13] C. Knapp, G. Carter, The generalized correlation method for estimation of time delay, IEEE Trans. on Acoustics, Speech, and Signal Processing 24 (4) (1976) 320–327.

[14] M. S. Brandstein, J. E. Adcock, H. F. Silverman, A closed-form location estimator for use with room environment microphone arrays, IEEE Trans. on Speech and Audio Processing 5 (1) (1997) 45–50.

[15] J. Benesty, J. Chen, Y. Huang, Time-delay estimation via linear interpolation and cross correlation, IEEE Trans. on Speech and Audio Processing 12 (5) (2004) 509–519.

[16] A. Karbasi, A. Sugiyama, A new doa estimation method using a circular microphone array, in: 2007 15th European Signal Processing Conference, IEEE, 2007, pp. 778–782.

[17] D. Bechler, K. Kroschel, et al., Considering the second peak in the gcc function for multi-source tdoa estimation with a microphone array, in: In Proceedings of the International Workshop on Acoustic Echo and Noise Control, Citeseer, 2003, pp. 315–318.

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

[19] 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.

[20] A. Saxena, A. Y. Ng, Learning sound location from a single microphone, in: 2009 IEEE International Conference on Robotics and Automation, IEEE, 2009, pp. 1737–1742.

[21] J. Chen, R. Takashima, X. Guo, Z. Zhang, X. Xu, T. Takiguchi, E. R. Hancock, Multimodal fusion for indoor sound source localization, Pattern Recognition 115 (2021) 107906.

[22] H. Do, H. F. Silverman, Srp-phat methods of locating simultaneous multiple talkers using a frame of microphone array data, in: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2010, pp. 125–128.

[23] W. Zhang, B. D. Rao, A two microphone-based approach for source localization of multiple speech sources, IEEE Trans. on Audio, Speech, and Language Processing 18 (8) (2010) 1913–1928.

[24] 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.

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

[26] R. Boora, S. K. Dhull, A tdoa-based multiple source localization using delay density maps, Sādhanā 45 (1) (2020) 1–12.

[27] H. Sundar, T. V. Sreenivas, C. S. Seelamantula, Tdoa-based multiple acoustic source localization without association ambiguity, IEEE/ACM Trans. on Audio, Speech, and Language Processing 26 (11) (2018) 1976–1990.

[28] R. Schmidt, Multiple emitter location and signal parameter estimation, IEEE Trans. on Antennas and Propagation 34 (3) (1986) 276–280.

[29] J. P. Dmochowski, J. Benesty, S. Affes, Broadband music: Opportunities and challenges for multiple source localization, in: 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, IEEE, 2007, pp. 18–21.

[30] C. T. Ishi, O. Chatot, H. Ishiguro, N. Hagita, Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environ-

ments, in: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2009, pp. 2027–2032.

[31] Y. Hu, T. D. Abhayapala, P. N. Samarasinghe, Multiple source direction of arrival estimations using relative sound pressure based music, IEEE/ACM Trans. on Audio, Speech, and Language Processing 29 (2020) 253–264.

[32] R. Roy, T. Kailath, Esprit-estimation of signal parameters via rotational invariance techniques, IEEE Trans. on acoustics, speech, and signal processing 37 (7) (1989) 984–995.

[33] L. O. Nunes, W. A. Martins, M. V. Lima, L. W. Biscainho, F. M. Gonçalves, A. Said, B. Lee, et al., A steered-response power algorithm employing hierarchical search for acoustic source localization using microphone arrays, IEEE Trans. on Signal Processing 62 (19) (2014) 5171–5183.

[34] M. Awad-Alla, A. Hamdy, F. A. Tolbah, M. A. Shahin, M. Abdelaziz, A two-stage approach for passive sound source localization based on the srp-phat algorithm, APSIPA Transactions on Signal and Information Processing 9.

[35] D.-B. Zhuo, H. Cao, Fast sound source localization based on srp-phat using density peaks clustering, Applied Sciences 11 (1) (2021) 445.

[36] F. Nesta, M. Omologo, Generalized state coherence transform for multidimensional tdoa estimation of multiple sources, IEEE Trans. on Audio, Speech, and Language Processing 20 (1) (2012) 246–260.

[37] N. Epain, C. T. Jin, Independent component analysis using spherical microphone arrays, Acta Acustica united with Acustica 98 (1) (2012) 91–102.

[38] T. Noohi, N. Epain, C. T. Jin, Direction of arrival estimation for spherical microphone arrays by combination of independent component analysis and sparse recovery, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 346–349.

[39] T. Noohi, N. Epain, C. T. Jin, Super-resolution acoustic imaging using sparse recovery with spatial priming, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 2414–2418.

[40] W. Cheng, Z. Zhang, G. Zhu, Z. He, Noise source identification and localization of mechanical systems based on an enhanced independent component analysis, Journal of Vibration and Control 22 (4) (2016) 1128–1142.

[41] J. Chen, J. Benesty, Y. A. Huang, Time delay estimation in room acoustic environments: an overview, EURASIP Journal on Advances in Signal Processing 2006 (1) (2006) 026503.

[42] J. H. DiBiase, A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays, Brown University Providence, RI, 2000.

[43] H. Do, H. F. Silverman, Y. Yu, A real-time srp-phat source location implementation using stochastic region contraction (src) on a large-aperture microphone array, in: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, Vol. 1, IEEE, 2007, pp. 121–124.

[44] J. Benesty, J. Chen, Y. Huang, Microphone array signal processing (2008).

[45] H. Sun, H. Teutsch, E. Mabande, W. Kellermann, Robust localization of multiple sources in reverberant environments using eb-esprit with spherical microphone arrays, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2011, pp. 117–120.

[46] H. Sun, E. Mabande, K. Kowalczyk, W. Kellermann, Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing, The Journal of the Acoustical Society of America 131 (4) (2012) 2828–2840.

[47] D. P. Jarrett, E. A. Habets, P. A. Naylor, 3d source localization in the spherical harmonic domain using a pseudointensity vector, in: 18th European Signal Processing Conference, IEEE, 2010, pp. 442–446.

[48] M. Swartling, B. Sällberg, N. Grbić, Source localization for multiple speech sources using low complexity non-parametric source separation and clustering, Signal Processing 91 (8) (2011) 1781–1788.

[49] D. Pavlidi, M. Puigt, A. Griffin, A. Mouchtaris, Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures, in: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2012, pp. 2625–2628.

[50] B. Loesch, B. Yang, Source number estimation and clustering for underdetermined blind source separation, in: in International Workshop on Acoustic Echo and Noise Control (IWAENC), Citeseer, 2008.

[51] S. Md, S. Goutam, Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition, Speech Communication 54 (2012) 543–565.

[52] D. T. Tran, E. Vincent, D. Jouvet, Extension of uncertainty propagation to dynamic mfccs for noise robust asr, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 5507–5511.

[53] B.-H. Juang, Maximum-likelihood estimation for mixture multivariate stochastic observations of markov chains, AT&T technical journal 64 (6) (1985) 1235–1249.

[54] T. Takiguchi, M. Nishimura, Y. Ariki, Acoustic model adaptation using first-order linear prediction for reverberant speech, IEICE Trans. INF. and SYST. E89-D (2006) 908–914.

[55] M. J. F. Gales, Predictive model-based compensation schemes for robust speech recognition, Speech Communication 25 (1998) 55–64.

[56] 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.

[57] M. Omologo, P. Svaizer, Use of the crosspower-spectrum phase in acoustic event location, IEEE Trans. on Speech and Audio Processing 5 (3) (1997) 288–292.

[58] 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) 1–9.