EM-based TDOA Estimation of a Speech Source via Gaussian Mixture Models in Noisy and Anechoic Environments

ZHIHUA LU1, JOÃO P. J. DA COSTA2, (Senior Member, IEEE), AND TAI FEI3, (Senior Member, IEEE)

1College of Information Science and Engineering, Ningbo University, Ningbo 315211, China (e-mail: luzhihua@nbu.edu.cn)
2Department 2 Lippstadt, Hamm-Lippstadt University of Applied Sciences (HSHL), Germany (e-mail: joaopaulo.dacosta@hshl.de)
3HELLA GmbH & Co. KGaA, 59552 Lippstadt, Germany (e-mail: tai.feil@ieee.org)

Corresponding author: Zhihua Lu (e-mail: luzhihua@nbu.edu.cn).

The research was supported by the National Natural Science Foundation of China (No. 61801255) and the K. C. Wong Magna Fund of Ningbo University.

ABSTRACT The propagation delay difference of a speech signal transmitted from the source to microphones, also known as time difference of arrival (TDOA), embodies the information of speech source position. The TDOA estimation plays a vital role in diverse systems such as teleconferencing and far-field speech recognition since the TDOA is a key parameter impacting quality of restored speech signals. This paper is devoted to estimating the TDOA of one speech source on a frame by frame basis in noisy and anechoic environments. First, we propose two variants of Gaussian mixture model to represent the speech signal received by a microphone pair, assuming Gaussianity of the signal and modeling speech sparsity by the speech presence probability (SPP). Second, after estimating the noise parameter in advance and formulating the speech parameters using the maximum likelihood principle, the proposed Gaussian mixture models are reduced to being dependent only on two unknowns, i.e. TDOA and SPP. Third, following these two models, we present two distinct estimators to estimate the TDOA and the SPP iteratively based on the expectation maximization algorithm. The proposed two estimators are free from the ad hoc parameter selection which is required in many classical approaches. Simulation results show that the TDOA estimated by them could be more accurate than that of the state-of-the-art GCC variants in a wide range of frames with specific SPP values. More importantly, the automatically estimated SPP which can be served as voice activity detection in a soft manner encodes the information of the TDOA estimation accuracy. In a speech frame, the estimated SPP with a large value indicates the estimated TDOA with small error. For example, when the SPP is larger than 0.76 and 0.87 in the two proposed estimators, respectively, the TDOA estimation error could be at most 19% of that in the worst case.

INDEX TERMS Expectation maximization, Gaussian mixture model, Speech presence probability, TDOA estimation

I. INTRODUCTION

Speech source localization, i.e., determining the spatial position of a speech source, is a fundamental issue in ad-hoc acoustic sensor networks composed of distributed microphones [1]–[3], and it finds a growing interest in many applications such as teleconferencing [4], far-field speech recognition [5], surveillance [6], and so on. The mainstream localization applications focus either on time difference of arrival (TDOA) estimation [7] [8] or direction of arrival (DOA) estimation [9]. Many source localization approaches have been developed in recent decades, which can be classified into two groups, i.e., spatial spectrum and time-frequency (TF) processing, see Table 1.

The spatial spectrum approaches construct a spectrum function of the spatial parameters (i.e., TDOA or DOA). The locations of highest peaks of the spectrum function indicate the TDOA (or DOA) candidates. The spectrum function can be constructed by methods like, e.g., the generalized cross-correlation (GCC) algorithm [10] and subspace-based methods. The GCC function is expressed by inserting a weight
into the cross-correlation function between the outputs of a microphone pair. The TDOA is estimated as the lag time that maximizes the GCC function. The most famous GCC function is the GCC with phase transform (GCC-PHAT). In [11], a non-linear function of the GCC is proposed in order to emphasize the large GCC values, which improves accuracy of the TDOA estimates. In [12], an empirical weight is proposed for the GCC function based on tuning two intrinsic parameters. The subspace-based methods rely on the eigen-structure of the array covariance matrix. The noise subspace or the signal subspace is used to construct the spectrum function which is likely to exhibit a large value at the true DOA. Two representatives are MUSIC [13] and ESPRIT [14]. Their variants suitable for wideband signals are proposed in [15] and [16], respectively. Note that the DOA estimate can be transferred to the TDOA estimate based on the microphone array geometry and planar wave assumption of speech signal. As shown in Table 1, the spatial spectrum approaches mentioned above do not make use of speech characteristics. The GCC variants can estimate the TDOA of a source using a single frame while the subspace-based approaches can deal with the case of multiple sources using a number of frames.

The TF processing approaches associate the TF bins to each source by means of a histogram [17] [18] or clustering [19] [20], and then use the associated TF bins to estimate the DOA (or TDOA) of each source, based on the assumption that at most one source is dominant in individual TF bins. The TF processing approaches have been investigated intensively in recent years due to its ability of jointly counting and localizing multiple sound sources even in the under-determined conditions. Different types of speech features are used to associate the TF bins to the unknown sources. Nevertheless, the TF processing approaches simultaneously analyze a large amount of measurements or snapshots (in other words, a large number of successive speech frames) and then the sources are localized in an offline manner with notable latency.

The maximum likelihood (ML) algorithm is a well-known tool for parameter estimation. Based on the assumption that the microphone outputs are Gaussian distributed, the ML algorithm is applied to estimate the TDOA [10], assuming that the parameters of the speech and noise are known a priori. The resulting function based on which the TDOA is estimated resembles the GCC function. Therefore, the ML-based TDOA estimator is simply treated as a member algorithm of the GCC family in the literature. In [21]–[24], the ML is used to iteratively estimate the DOA (or TDOA) as well as the parameters of the speech and noise based on a large number of successive speech frames. The estimation procedures follow the principle of the TF processing approaches (more precisely, clustering), see Table 1.

Recently, deep learning is applied to estimate the TDOA, which is shown to be more robust in adverse environments. In [25], deep neural networks (DNNs) are applied to predict a time-frequency mask which is combined with the GCC-PHAT to estimate the TDOA. In [26], convolutional neural networks (CNNs) are devised to jointly fulfill the tasks of denoising and dereverberation, before the TDOA is estimated by the GCC. In [27], an end-to-end DNN based acoustic source localization approach is proposed. However, when a neural network that is trained by one data set is applied to another data set from a different distribution, its performance is not guaranteed [28].

As shown in Table 1, none of them can incorporate speech characteristics while estimating the TDOA in real time (using a single frame). We attempt to fill the gap in this paper. In order to fulfill the practical requirement of real-time update of TDOA estimates, we estimate the TDOA based on one frame of speech signals, and update the TDOA estimates frame by frame. To incorporate speech characteristics, we use the speech presence probability (SPP) [29] [30] to model speech sparsity and study the role of the SPP in the TDOA estimation.

Three assumptions are made for speech signals received by a microphone pair. First, the noise is a Gaussian random variable. Second, the speech is sparse, in other words, the speech could be absent at some times and frequencies. Third, when the speech is present, it could be treated as either a Gaussian random variable or a deterministic value. It results in two models for the noisy speech signals, which contain a number of unknown parameters of the speech and noise. The noise parameter (i.e., variance) is assumed to be known. Otherwise, it is estimated in advance. Underlying the principle of the ML algorithm, the speech parameters can be formulated in terms of the received signals and the mixing vector of

---

**Table 1. Summary of the TDOA (DOA) estimation approaches**

| Approach       | Reference | Frame size | Speech characteristics | Source number |
|----------------|-----------|------------|------------------------|---------------|
| Spatial spectrum | GCC variants | Single     | No                     | 1             |
|                | Subspace-based | Multiple   | No                     | ≥1            |
| T-F processing | Histogram | Multiple   | Yes                    | ≥1            |
|                | Clustering | Multiple   | Yes                    | ≥1            |
| Deep learning  | [25]–[28] | Multiple   | Yes                    | ≥1            |
| Hybird         | Proposed  | Single     | Yes                    | 1             |

---

1 In the case of multiple sources, their DOA (or TDOA) can be estimated after implementing various blind source separation algorithms. This issue is beyond the scope of the paper.
the microphone pair as well as the noise variance. If these formulations are plugged back into the two models of noisy speech signals, the resulting two Gaussian mixture models (GMMs) will only depend on the TDOA (contained in the mixing vector) and the SPP. Then, the TDOA and the SPP are estimated iteratively using the expectation maximization (EM) algorithm [31]. Performance improvement of the proposed two approaches over some existing GCC approaches is observed in our simulations. Moreover, in contrast to the state of the art approaches, the proposed approaches automatically estimate the SPP for the audio recordings, avoiding the usage of voice activity detection in case that the recordings should be further processed. The estimated SPP is able to tell how accurate the estimated TDOA is. Therefore, the proposed approaches complete two tasks simultaneously: estimate the TDOA and also give the estimation accuracy.

The proposed approaches are categorized into hybrid approach in Table 1 since the proposed approaches can be elaborated from two different viewpoints. First, the TDOA is estimated using the spatial spectrum that is constructed based on the likelihood function. Second, the EM-based estimation procedure using a single frame includes a clustering step in the frequency domain. More precisely, in E-step, the frequency bins are divided softly into two clusters, i.e., noisy speech and pure noise, before the TDOA is estimated as an unknown parameter of the cluster of noisy speech in M-step. In contrast, the clustering using multiple frames in [21]–[24] takes place in time-frequency domain, where all clusters correspond to speech sources and no cluster deals with pure noise.

It is noteworthy that when super-Gaussian distributions are used to model speech signals instead of Gaussian distributions, the speech enhancement performance is improved [32]. Due to the fact that super-Gaussian distributions are more appropriate to represent speech signals, it is reasonable to believe that the TDOA estimation accuracy could be improved if super-Gaussian distributions are used in our proposed approaches. An extension of our proposed approaches to super-Gaussian cases will be done in the future work.

The remainder of the paper is organized as follows. In Section II, the problem of TDOA estimation is formulated briefly. In Section III, two GMMs for the noisy speech signals are introduced. In Section IV, the EM-based TDOA estimation procedure is presented and discussed. In Section V, simulation results and performance analysis are given. Finally, conclusions are drawn in Section VI.

II. PROBLEM FORMULATION

Consider the case where a single speech signal emanating from a source in the far-field is recorded by two spatially separated microphones in an anechoic and noisy environment. Let \( x_1(t) \) and \( x_2(t) \) denote the signals received at the microphone pair. Since there is only the line-of-sight (LOS) component due to the anechoic assumption, without loss of generality, the signals \( x_1(t) \) and \( x_2(t) \) can be modeled as

\[
\begin{align}
x_1(t) &= s(t) + n_1(t) \\
x_2(t) &= s(t - \tau) + n_2(t),
\end{align}
\]

where \( t \) is discrete time, \( s(t) \) is the source speech signal, \( n_1(t) \) and \( n_2(t) \) are the noise. In \( x_2(t) \), \( \tau \) denotes the relative time delay, which is known as the TDOA of the speech signal transmitted from the source to the microphone pair.

The speech signals in (1) are segmented into frames. Transforming the signals of the \( l \)th frame using the short-time Fourier transform (STFT), we derive

\[
\begin{align}
X_1(l, f) &= S(l, f) + N_1(l, f) \\
X_2(l, f) &= e^{-j2\pi f \tau}S(l, f) + N_2(l, f),
\end{align}
\]

where \( f \) is the frequency bin. In the sequel, the frame index \( l \) will be omitted for brevity when there is no confusion. We can rewrite (2) in a vector form as

\[
X(f) = A(f)S(f) + N(f),
\]

where

\[
\begin{align}
X(f) &= [X_1(f), X_2(f)]^T \\
A(f) &= [1, e^{-j2\pi f \tau}]^T \\
N(f) &= [N_1(f), N_2(f)]^T,
\end{align}
\]

with \([.]^T\) standing for matrix transpose and \( A(f) \) denoting the mixing vector.

Our goal is to find \( \tau \) given only the noisy measurements (i.e., the output of the microphone pair). In Section III, we present how the noisy speech signals are modeled by GMM, while, in Section IV, an EM-based procedure is proposed to solve the TDOA estimation problem.

III. MODEL OF THE NOISE AND SPEECH

We assume that \( n_1(t) \) and \( n_2(t) \) in (1) are Gaussian distributed and uncorrelated with each other. Under mild conditions, the Fourier coefficients of the noise, i.e., \( N_1(f) \) and \( N_2(f) \) in (2), can be treated as statistically independent complex Gaussian random processes. Then, we have the probability density function (pdf) of \( N(f) \) given by

\[
p(N(f)) = \frac{1}{\pi^2 \det \Sigma(f)} \exp \left[ -N^*(f)\Sigma(f)^{-1}N(f) \right]
\]

with the covariance matrix

\[
\Sigma(f) = \begin{bmatrix}
\sigma^2_{n1}(f) & 0 \\
0 & \sigma^2_{n2}(f)
\end{bmatrix},
\]

where \( \sigma^2_{n1}(f) \) and \( \sigma^2_{n2}(f) \) are the variance of \( N_1(f) \) and \( N_2(f) \), respectively. Note that they are estimated prior to the procedure of TDOA estimation. Herein, \( \det[.] \) stands for the determinant of a matrix and \([.]^*\) stands for the complex conjugate transpose.

When the speech is absent, i.e., \( S(f) = 0 \), we have \( X(f) = N(f) \) according to (3). Then, by replacing \( N(f) \) with \( X(f) \) in (5), we derive the pdf of \( X(f) \) as follows,

\[
g_0(X(f)) = \frac{1}{\pi^2 \det \Sigma(f)} \exp \left[ -X^*(f)\Sigma(f)^{-1}X(f) \right],
\]

(7)
which is a two-dimensional complex Gaussian pdf with zero-mean and the covariance \( \Sigma(f) \) given in (6).

When the speech is present, namely \( S(f) \) is non-zero, the pdf of \( X(f) \) is then denoted by \( g_1(X(f)) \). To take two cases that the speech is absent and present into account, we formulate the pdf of \( X(f) \) in (3) using the Gaussian mixture model (GMM):

\[
p(X(f)) = p_0g_0(X(f)) + p_1g_1(X(f)), \tag{8}
\]

where \( p_0 \) and \( p_1 \) denote the a priori speech absence probability (SAP) and speech presence probability (SPP) in a single frame, respectively. It is clear that \( p_0 + p_1 = 1 \).

When the speech is present, \( S(f) \) can be modeled as either a Gaussian random process or a deterministic arbitrary value, according to array signal processing [33]. Accordingly, we can model the pdf of \( X(f) \) using the stochastic GMM (GMM-s) or the deterministic GMM (GMM-d) as shown in Subsections III-A and III-B, respectively.

### A. STOCHASTIC GAUSSIAN MIXTURE MODEL (GMM-S)

In (3), the noise \( N(f) \) is assumed to be a zero-mean Gaussian random process, see (5). Now, the speech signal \( S(f) \) is also treated as a Gaussian random process with zero-mean. According to the model in (3), \( X(f) \) is a Gaussian random variable with the pdf

\[
g_{1s}(X(f)) = \frac{1}{\pi^2 \det \Gamma(f)} \exp \left\{ -X^*(f)\Gamma(f)^{-1}X(f) \right\}, \tag{9}
\]

which is a two-dimensional complex Gaussian pdf with zero-mean and the covariance matrix

\[
\Gamma(f) = A(f)A^*(f)\sigma_s^2(f) + \Sigma(f), \tag{10}
\]

where \( A(f) \) (see (4)) is the mixing vector, and \( \Sigma(f) \) (see (6)) is the covariance matrix of the noise \( N(f) \), and \( \sigma_s^2(f) \) denotes the variance of the speech \( S(f) \). \( \sigma_s^2(f) \) is the only unknown parameter of the speech in this case can be estimated by the ML algorithm as follows [33] [34]:

\[
\hat{\sigma}_s^2(f) = B(f)[X(f)X^*(f) - \Sigma(f)]B^*(f), \tag{11}
\]

with

\[
B(f) = \frac{A^*(f)\Sigma(f)^{-1}}{A^*(f)\Sigma(f)^{-1}A(f)}. \tag{12}
\]

Here, \( B(f) \) is involved to represent a matrix consisting of \( A(f) \) and \( \Sigma(f) \), for ease of notation.

Denote \( \hat{g}_{1s}(X(f)) \) by the pdf in (9) after \( \sigma_s^2(f) \) is substituted by the estimate \( \hat{\sigma}_s^2(f) \) in (11). Since \( \Sigma(f) \) is estimated in advance, the pdf \( \hat{g}_{1s}(X(f)) \) possesses only one unknown parameter, that is, the TDOA (in the mixing vector \( A(f) \)).

Substituting \( \hat{g}_{1s}(X(f)) \) for \( g_1(X(f)) \) in (8), we obtain the new pdf as follows:

\[
p_s(X(f)) = p_0g_0(X(f)) + p_1\hat{g}_{1s}(X(f)), \tag{13}
\]

which is referred to as the GMM-s in this paper.

### B. DETERMINISTIC GAUSSIAN MIXTURE MODEL (GMM-D)

If \( S(f) \) in (3) is treated as a deterministic arbitrary value, we obtain the pdf \( g_{1d}(X(f)) \) as follows:

\[
g_{1d}(X(f)) = \frac{1}{\pi^2 \det \Sigma(f)} \exp \left\{ -\frac{1}{2} \left[ X(f) - A(f)S(f) \right]^T \Sigma(f)^{-1} \left[ X(f) - A(f)S(f) \right] \right\}, \tag{14}
\]

which is a two-dimensional complex Gaussian pdf with the mean \( A(f)S(f) \) and the covariance matrix \( \Sigma(f) \). In (14), \( A(f) \) and \( \Sigma(f) \) are given in (4) and (6), respectively. As the unknown parameter of the speech, \( S(f) \) can be estimated by the ML algorithm as follows [33] [35]:

\[
\hat{S}(f) = B(f)X(f), \tag{15}
\]

where \( B(f) \) is a matrix consisting of \( A(f) \) and \( \Sigma(f) \), given in (12).

Denote by \( \hat{g}_{1d}(X(f)) \) the pdf in (14) after \( S(f) \) being substituted by the estimate \( \hat{S}(f) \) in (15). The unknown parameter of \( \hat{g}_{1d}(X(f)) \) consists of only the TDOA (in \( A(f) \)), the same as that of \( \hat{g}_{1s}(X(f)) \).

Substituting \( \hat{g}_{1d}(X(f)) \) for \( g_1(X(f)) \) in (8), we obtain the new pdf as follows:

\[
p_d(X(f)) = p_0g_0(X(f)) + p_1\hat{g}_{1d}(X(f)), \tag{16}
\]

which is referred to as the GMM-d in this paper.

### IV. EM-BASED TDOA ESTIMATION ALGORITHM

#### A. IMPLEMENTATION OF THE ALGORITHM

Both the GMM-s in (13) and the GMM-d in (16) possess two unknown parameters, namely the a priori SPP \( p_1 (p_0 = p_1 - 1) \) and the TDOA \( \tau \), based on the fact that the noise variance is estimated in advance. We can estimate these unknown parameters using the EM algorithm. In the treatment of EM algorithm, \( p_1 \) (equivalently \( p_0 \)) is considered as a hidden variable, while \( \tau \) is considered as an unknown variable of the model. Based on the general form of \( p(X(f)) \) in (8), the EM estimation using a speech frame contains two iterative steps:

- **E-step:**

  \[
  p_{k|X(f)}^{(i)} = \frac{p_{k|X(f)}^{(i-1)}g_k(X(f); \tau^{i-1})}{p_0^{(i-1)}g_0(X(f); \tau^{i-1}) + p_1^{(i-1)}g_1(X(f); \tau^{i-1})}, \quad k = 0, 1 \tag{17}
  \]

- **M-step:**

  \[
  p_k^{(i)} = \frac{1}{F} \sum_f p_{k|X(f)}^{(i)} \tag{18}
  \]

\[
\tau^{(i)} = \arg \max_{\tau} \sum_f p_{1|X(f)}^{(i)} \log [g_1(X(f); \tau)], \tag{19}
\]

where \( i \) is the iteration index, \( F \) denotes the number of frequency bins, \( p_{k|X(f)}^{(i)} \) denotes the estimate of the conditional
SAP (or SPP) in the \(i\)th iteration given \(X(f)\) (or simply the a posteriori SAP (or SPP)), \(p_{1}^{(i)}\) denotes the estimate of the a priori SAP (or SPP) in the \(i\)th iteration, and \(\tau^{(i)}\) denotes the TDOA estimate in the \(i\)th iteration. Note that the EM iteration starts with initial estimates (i.e., \(p_{0}^{(0)}, \tau^{(0)}\)) and stops when the estimation is converged. The EM-based estimation algorithm in each frame is summarized in Table 2. More implementation details will be given in Section V-A.

### TABLE 2. EM-based TDOA estimation algorithm

1. Pre-estimate \(\Sigma(f)\) (see (6)) using the traditional algorithms (e.g., [36]).
2. Set \(i = 1\), initialize \(p_{0}^{(0)}\) and \(\tau^{(0)}\).
3. While not converged and \(i \leq \) no. of iters do
   4. Update \(g_{0}(X(f), \tau^{(i-1)})\) using (7) for all frequency bins.
   5. Update \(g_{1}(X(f), \tau^{(i-1)})\) using (9) or (14) for all frequency bins.
   6. Update \(p_{1}^{(i)}\) using (17) for all frequency bins.
   7. Update \(\tau^{(i)}\) using (19).
   8. \(i = i + 1\).
4. End while

If we replace the pdf in (8) with the GMM-s in (13) or the GMM-d in (16) in the EM iteration, more precisely, substituting \(\hat{g}_{1s}(X(f))\) or \(\hat{g}_{1d}(X(f))\) for \(g_{1}(X(f))\) in (17) and (19), we can derive two types of TDOA estimates, i.e., \(\tau_{s}\) and \(\tau_{d}\), respectively, after the EM iteration stops. Also, we have two different a priori SPPs, that is, \(p_{1s}\) and \(p_{1d}\).

### B. DISCUSSION ABOUT THE ALGORITHM

The EM-based estimation procedure of the GMMs can be elaborated in the viewpoint of soft clustering in the frequency domain. In E-step (see (17)), the a posteriori SAP (or SPP) \(p_{k}^{(i)}(X(f))\) with \(k = 0, 1\) is calculated for each frequency bin of a frame, which accounts for the probability of the frequency bin to be in two clusters, i.e., pure noise and noisy speech. It means that all frequency bins are divided softly into two clusters. In M-step (see (18) and (19)), the unknown parameters of the clusters are estimated, including the a priori SAP (or SPP) \(p_{k}^{(i)}\) and the TDOA.

If hard clustering is applied in our problem, each frequency bin either belongs to a cluster completely or not. Since the noise pdf in (7) does not depend on the TDOA, the TDOA should be estimated using only the frequency bins belonging to the cluster of noisy speech. Similarly, for the TDOA estimation in soft clustering, more emphasis is placed on the frequency bins with a high probability to be in the cluster of noisy speech (i.e., large \(p_{1}^{(i)}\)). Details are going to be explained in the following.

In (19), the term to be maximized on the right side of the equation can be considered as the spatial spectrum (see Section I), which is equivalent to a weighted sum of the log-likelihood functions of all frequency bins, where the weight is the a posteriori SPP \(p_{1}^{(i)}\), and the log-likelihood function is derived from the pdf of noisy speech (see (9) or (14)). Note that the a posteriori SPP \(p_{1}^{(i)}\) accounts for the probability of a frequency bin to be in the cluster of noisy speech. The frequency bins with a high probability to be in the cluster of noisy speech will be emphasized in the spatial spectrum. Meanwhile, the impact of remaining frequency bins on the final spatial spectrum is marginal since their a posteriori SPPs are low.

In (18), the a priori SPP \(p_{1}^{(i)}\) is computed as the average of the a posteriori SPPs \(p_{1}^{(i)}(X(f))\). Therefore, in a frame with a relatively large \(p_{1}^{(i)}\), the majority of frequency bins are expected to have high probabilities to be in the cluster of noisy speech. Simply speaking, \(p_{1}^{(i)}\) can be served as soft voice activity detection in a frame. Its value indicates frame quality, namely, the percentage of frequency bins that is much more likely to be in the cluster of noisy speech. In a frame of good quality, the TDOA estimated by the proposed approaches is relatively accurate.

As aforementioned, \(g_{0}(X(f))\) does not contain any unknown parameters whereas both \(\hat{g}_{1s}(X(f))\) and \(\hat{g}_{1d}(X(f))\) possess the TDOA as the unknown parameter. When the TDOA is determined, the variance of \(\hat{g}_{1s}(X(f))\) (with zero mean) and the mean of \(\hat{g}_{1d}(X(f))\) (with the same variance as \(g_{0}(X(f))\)) increases as \(|X(f)|\) increases. Here, \(|X(f)|\) is the norm of \(X(f)\). At a high SNR level, as \(|X(f)|\) increases (i.e., the spectral amplitude of the speech increases), the values of \(\hat{g}_{1s}(X(f))\) and \(\hat{g}_{1d}(X(f))\) gradually overwhelm that of \(g_{0}(X(f))\). Therefore, the a posteriori SPP in (17) is apt to be overestimated. As a result, the a priori SPP in (18) which is the mean of the a posteriori SPP is also overestimated. It will be observed in the simulation results.

The pdf \(\hat{g}_{1s}(X(f))\) can be formulated as the Gaussian distribution:

\[
N_{c}(0, A(f)B(f)X(f)X^{*}(f)B^{*}(f)A^{*}(f) + \beta),
\]

where \(\beta\) denotes the term independent of \(X(f)\). The pdf \(\hat{g}_{1d}(X(f))\) can be formulated as the Gaussian distribution:

\[
N_{c}(A(f)B(f)X(f), \Sigma(f)).
\]

By comparing the above two Gaussian distributions, we can derive that \(\hat{g}_{1s}(X(f)) < \hat{g}_{1d}(X(f))\) in case that \(|X(f)|\) is large enough. Therefore, we have \(p_{1s} < p_{1d}\), according to (17) and (18).

### V. SIMULATION RESULTS

#### A. EXPERIMENTAL SETUP

The ISM MATLAB code [37] is used to generate anechoic acoustic response impulses. In a room with the size of 3 × 4 × 2.5 meter, two microphones are placed with 4 possible distances between each, that is, 5 cm, 10 cm, 15 cm and 20 cm. For each location of the microphone pair, the speech source is placed at different positions. It results in four TDOA values, that is, 7.0691 × 10^{-4} s, -1.4126 × 10^{-4} s, 0 s and 6.7436 × 10^{-4} s. We randomly select 8 sentences from the TIMIT database for the source speech signal. From the Noisex-92 database, 10 samples of uncorrelated two-channel noise are constructed, including white noise, babble noise and factory noise. The noise variance is estimated in advance using well-known noise estimation algorithms, e.g., Martin’s
minimum tracking algorithm [36]. The sampling rate is set to 16 kHz. The speech signal is segmented into frames of length 512, and successive frames are overlapped by 50%. For the STFT, a Hanning window of length 512 is used. The TDOA is estimated for each frame.

We select three candidate estimators: GCC-PHAT [10], GCC-NONLIN [11] and \( \rho \)-PHAT-C [12], which are suitable for real-time estimation using a single frame. For the \( \rho \)-PHAT-C, two tuning parameters are empirically recommended in [12], that is, \( \rho = 1 \) and \( b = 0.75 \). EM-GMM-s and EM-GMM-d are the abbreviations for the two proposed estimators based on the GMM-s and GMM-d, respectively. The EM iteration is initialized as follows: \( p_1(0) = 0.5 \) and \( \tau(0) \) equals the TDOA estimate of the GCC-NONLIN. We set the maximal iteration number as 2 due to the fact that it yields satisfactory estimation results in our simulations. In addition, maximization in (19) is done by grid search on the TDOA candidates with the resolution of \( \frac{1}{1000} \) [21]. Here, \( d \) and \( c \) denote the distance between the microphone pair and the propagation velocity of the speech signal, respectively.

### B. WHITE NOISE AND ANECHOIC ENVIRONMENTS

In this subsection, the white noise and anechoic environments are considered. Based on the settings above, we have 320 samples of noisy two-channel speech signals at each of three SNR levels (i.e., 5 dB, 10 dB and 20 dB). Table 3 shows the root mean squared error (RMSE) and the mean absolute error (MAE) of the TDOAs estimated by five estimators for all the 60680 speech frames segmented from 320 speech samples at each SNR level. We can find that the MAE coincides with the RMSE in terms of performance comparison. The EM-GMM-s and EM-GMM-d outperform the GCC-PHAT and GCC-NONLIN. However, they are slightly defeated by the \( \rho \)-PHAT-C. It should be noted that the \( \rho \)-PHAT-C suffers from a lack of theoretical justification since its two parameters are selected empirically. In contrast, the proposed two estimators are free from any ad hoc parameter selection.

| Error type | SNR (dB) | 5  | 10  | 20  |
|------------|---------|----|-----|-----|
| RMSE       |         |    |     |     |
| GCC-PHAT   | 0.4287  | 0.3714 | 0.2944 |
| GCC-NONLIN | 0.4516  | 0.3587 | 0.2851 |
| \( \rho \)-PHAT-C | 0.3636 | 0.3229 | 0.2663 |
| EM-GMM-s   | 0.3700  | 0.3300 | 0.2780 |
| EM-GMM-d   | 0.3679  | 0.3245 | 0.2700 |
| MAE        |         |    |     |     |
| GCC-PHAT   | 0.2345  | 0.1812 | 0.1125 |
| GCC-NONLIN | 0.2163  | 0.1644 | 0.1028 |
| \( \rho \)-PHAT-C | 0.1724 | 0.1342 | 0.0888 |
| EM-GMM-s   | 0.1775  | 0.1389 | 0.0935 |
| EM-GMM-d   | 0.1751  | 0.1353 | 0.0904 |

In addition to the TDOA, the EM-GMM-s and EM-GMM-d also estimate the a priori SPP in (18) (i.e., \( p_{1s} \) and \( p_{1d} \), respectively) in each frame, which contains valuable information. In what follows, we investigate the impact of the a priori SPPs on the estimated TDOAs in detail. Note that the SPP means the a priori SPP except where noted.

Figs. 1-2 show the relationship between the TDOAs and the SPPs estimated by the proposed two estimators for the frames of a noisy speech sample. In Fig. 1, the solid line represents the true TDOA value and the mark ‘+’ represents a pair of TDOA and SPP \((p_{1s})\) estimated by the EM-GMM-s in a frame. In the frames with \( p_{1s} \) larger than a certain value, e.g., 0.75, the marks ‘+’ are located much closer to the solid line, which means that the estimation error of the TDOA is significantly reduced, as compared with that in the frames with \( p_{1s} < 0.75 \). In Fig. 2, similar results can be observed when \( p_{1d} \) is larger than 0.87. In the frames where the speech is present with a relatively high probability (i.e., large \( p_{1s} \) or \( p_{1d} \)), the frequency bins are more likely to contain noisy speech components, which indicates relatively good frame quality and consequently leads to the TDOA estimates with small error. This has been explained in Section IV-B.

We rank the SPPs calculated in all the 60680 frames in a descending order and partition the SPPs into different levels. As a result, the frames with different levels of quality are selected according the SPP levels. For example, the frames with top 10% SPPs are exactly the top 10% frames of best quality. Figs. 3-4 show the thresholds of the SPPs when different percentage (from 99% to 10%) of the largest SPP values is selected. It is found that the threshold of \( p_{1s} \) ranges from 0.618 to 0.891 whereas that of \( p_{1d} \) ranges from 0.698 to 0.952, considering all the three SNR levels. Additionally, Figs. 3-4 tell us how the estimated \( p_{1s} \) and \( p_{1d} \) are roughly distributed in different conditions.

The lower limit of \( p_{1s} \) and \( p_{1d} \) is overestimated since the theoretical value should be close to 0.5.2 Also, the estimated \( p_{1s} \) is smaller than the counterpart \( p_{1d} \). This is due to the

\[^2\]In a frame with speech absence, theoretically, \( g_1(X(f)) \) in (8) could degenerate to \( g_0(X(f)) \). These two pdfs are indistinguishable, which results in that the a priori SPP (i.e., \( p_1 \)) equals 0.5.
formulation difference between $\hat{g}_{1s}(X(f))$ and $\hat{g}_{1d}(X(f))$, which has been elaborated at the end of Section IV-B.

Figs. 5-7 show the RMSE of the TDOA estimates in the frames at different SPP levels. $p_{1s}$ is used to select the frames for the EM-GMM-s, the GCC-NONLIN and the $\rho$-PHAT-C, whereas $p_{1d}$ is used for the EM-GMM-d. The RMSE of the TDOA values estimated by both the EM-GMM-s and EM-GMM-d decreases in the frames with increased SPPs. In other words, the proposed two estimators tend to estimate the TDOA more accurately in a frame with a larger SPP. The RMSE curves of the EM-GMM-s in Figs. 5-7 descend faster than the ones of the EM-GMM-d. The EM-GMM-s gives smaller RMSE values than the EM-GMM-d in the frames with large SPPs. Hence, $p_{1s}$ is more competent than $p_{1d}$ in indicating the TDOA estimation accuracy. It means that the EM-GMM-s is more precise than the EM-GMM-d in terms of modeling noisy speech signals in the frames of relatively good quality (i.e., relatively large SPPs). Additionally, the EM-GMM-s and the $\rho$-PHAT-C perform similarly in the frames with relatively large SPPs. However, the former is slightly worse than the latter in the frames with small SPPs. This might be caused by the fact that the EM performance is degraded when the EM iteration converges to the local optima in the frame with bad quality (i.e., small SPPs).

By combing the results shown in Figs. 5-7 with those in Table 4, the proposed two estimators tend to estimate the TDOA more accurately in a frame with a larger SPP. The RMSE curves of the EM-GMM-s in Figs. 5-7 descend faster than the ones of the EM-GMM-d. The EM-GMM-s gives smaller RMSE values than the EM-GMM-d in the frames with large SPPs. Hence, $p_{1s}$ is more competent than $p_{1d}$ in indicating the TDOA estimation accuracy. It means that the EM-GMM-s is more precise than the EM-GMM-d in terms of modeling noisy speech signals in the frames of relatively good quality (i.e., relatively large SPPs). Additionally, the EM-GMM-s and the $\rho$-PHAT-C perform similarly in the frames with relatively large SPPs. However, the former is slightly worse than the latter in the frames with small SPPs. This might be caused by the fact that the EM performance is degraded when the EM iteration converges to the local optima in the frame with bad quality (i.e., small SPPs).
Figs. 3-4, we can find that both $p_{1s}$ and $p_{1d}$ can serve as the indicators to tell accuracy of the estimated TDOA values of the proposed estimators. If the estimators are expected to achieve a certain RMSE level of TDOA estimates, we have to select the frames of sufficiently good quality, namely the percentages of the largest SPPs at individual SNR in Figs. 5-7, respectively. Furthermore, according to this number, we can find the corresponding threshold of the SPPs in Figs. 3-4. Table 4 shows the percentages and thresholds of $p_{1s}$ and $p_{1d}$ at three RMSE levels, i.e., $1.5 \times 10^{-4}$, $1 \times 10^{-4}$ and $0.5 \times 10^{-4}$. It is found that the percentages are fairly different in different conditions. However, the thresholds for $p_{1s}$ and $p_{1d}$ are located in small regions, i.e., $[0.73, 0.76]$ and $[0.84, 0.87]$, respectively. If the estimated $p_{1s}$ and $p_{1d}$ are located in those two regions (indicating relatively good frame quality), the RMSE of the TDOA estimates is relatively small, e.g., at most $1.5 \times 10^{-4}$ of the largest RMSE at 20 dB SNR and $0.5 \times 10^{-4}$ of the largest RMSE at 5 dB SNR. Therefore, when $p_{1s} > 0.76$ or $p_{1d} > 0.87$, the RMSE of the TDOA estimates is supposed to be at most 19% of that in the worst case, irrespective of SNR levels. We state that we can roughly predict accuracy of the estimated TDOA value according to the estimated SPP value in each frame. In other words, the proposed two estimators not only estimate the TDOA but also indicate accuracy of the estimation.

**C. NON-WHITE NOISE AND REVERBERANT ENVIRONMENTS**

Although the proposed estimators are originally designed for white noise and anechoic environments, their performance in non-white noise and reverberant environments is also studied. Figs. 8 and 9 show the RMSE of the TDOA estimates at different SPP levels in babble noise and anechoic environments at 5 dB and 20 dB SNR, respectively. Figs. 10 and 11 show the RMSE of the TDOA estimates in factory noise and anechoic environments at 5 dB and 20 dB SNR, respectively. In both babble and factory noise, the relationship between the RMSE of TDOA estimates and the SPPs remains the same as that in white noise, i.e., the RMSE decreases as the SPP increases. The SPP estimated in the EM-GMM-s outperforms the counterpart in the EM-GMM-d in terms of indicating the TDOA estimation accuracy, since the RMSE curves of the EM-GMM-s in Figs. 8-11 descend faster than the ones of the EM-GMM-d. The performance of the $\rho$-PHAT-C is degraded severely in non-white noise due to the fact that its performance is highly dependent on the two tuning parameters which are selected without theoretical justification.

Figs. 12 and 13 show the RMSE of the TDOA estimates in white noise (10 dB SNR) and reverberant environments with T60 = 0.1 s and 0.2 s $^3$, respectively. The RMSE curves of the two proposed estimators in Fig. 12 descend as the SPP rises except for the 10% of the largest SPPs. The SPP still can roughly indicate the TDOA estimation accuracy when T60 = 0.1 s. However, in Fig. 13, the RMSE of TDOA estimates is not inversely proportional to the SPP. The SPP is not able to indicate the TDOA estimation accuracy when T60 = 0.2 s. This can be explained in the following. The proposed GMMs group the speech signals into two clusters, namely noisy speech and pure noise. The reverberant speech will destroy the validity of the GMMs and prevent the EM algorithm from accurately identifying two clusters. The strong part of the reverberant speech is apt to be identified to be the cluster of noisy speech. For example, a frequency bin of a frame does not contain speech signals from direct path but contain the reverberant speech of previous frames. This frequency bin could be identified to be the cluster of noisy speech.

$^3$T60 accounts for the time it takes for sound to decay by 60 dB in a room or space.
Therefore, the estimated SPP becomes inaccurate, especially when the reverberant condition becomes severe. In order to cope with the reverberant condition, the proposed GMMs and EM estimation procedure should take the dereverberation step into account. This is expected to be studied in the future.

![Figure 8](image8.png)

**FIGURE 8.** RMSE of TDOA estimates in the frames with different SPP levels at 5 dB SNR in babble noise environments.

![Figure 9](image9.png)

**FIGURE 9.** RMSE of TDOA estimates in the frames with different SPP levels at 20 dB SNR in babble noise environments.

![Figure 10](image10.png)

**FIGURE 10.** RMSE of TDOA estimates in the frames with different SPP levels at 5 dB SNR in factory noise environments.

![Figure 11](image11.png)

**FIGURE 11.** RMSE of TDOA estimates in the frames with different SPP levels at 20 dB SNR in factory noise environments.

VI. CONCLUSIONS

In this paper, we propose an EM-based TDOA estimation procedure via stochastic and deterministic GMM for single speech in noisy and anechoic environments. By carefully modeling the speech and the noise, the two GMMs are independent of the speech parameters, and the only parameter to be pre-estimated is the noise variance. After that, the TDOA and SPP values are estimated iteratively via the EM algorithm. Simulations show that the proposed estimators outperform or are comparable to the selected state-of-the-art estimators in terms of TDOA estimation. Moreover, the a priori SPPs calculated in the proposed estimators which can be served as soft voice activity detection, are utilized as the indicators for accuracy of TDOA estimates. Therefore, the proposed estimators complete two tasks simultaneously: estimate the TDOA and also give the estimation accuracy.

REFERENCES

[1] J. Yang, X. Zhong, W. Chen, and W. Wang, “Multiple acoustic source localization in microphone array networks,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 29, pp. 334–347, 2021.
[2] A. Brendel and W. Kellermann, “Distributed source localization in acoustic sensor networks using the coherent-to-diffuse power ratio,” IEEE J. Sel. Topics Signal Process., vol. 13, no. 1, pp. 61–75, 2019.
[3] D. Paredes and J. Apolinario Jr., “Shooter localization based on DoA estimation of gunshot signals and digital map information,” IEEE Lat. Am. Trans., vol. 13, no. 2, pp. 441–447, Feb. 2015.
[4] V. A. Nguyen and et al., “ITEM: immersive telepresence for entertainment and meetings - a practical approach,” IEEE J. Sel. Topics Signal Process., vol. 9, no. 3, pp. 546–561, 2015.
[5] A. S. Subramanian and et al., “Directional ASR: a new paradigm for E2E
FIGURE 12. RMSE of TDOA estimates in the frames with different SPP levels at T60 = 0.1 s in reverberant environments.

FIGURE 13. RMSE of TDOA estimates in the frames with different SPP levels at T60 = 0.2 s in reverberant environments.

multi-speaker speech recognition with source localization,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2021, pp. 8433–8437.

[6] B. Chen, C. Chen, and J. Wang, “Smart homecare surveillance system: Behavior identification based on state-transition support vector machines and sound directivity pattern analysis,” IEEE Trans. Syst. Man Cybern. Syst., vol. 43, no. 6, pp. 1279–1289, 2013.

[7] J. Chen, J. Benesty, and Y. Huang, “Time delay estimation in room acoustic environments: An overview,” EURASIP J. Adv. Signal Process., vol. 2006, pp. 1–19, Article ID 26503, 2006.

[8] Y. Huang, J. Benesty, and J. Chen, “Time delay estimation and source localization,” in Springer Handbook of Speech Processing, J. Benesty, M. Sondhi, and Y. Huang, Eds. Springer, Berlin, Heidelberg, 2008.

[9] R. Zekavat and R. M. Buehrer, “Source localization: Algorithms and analysis,” in Handbook of Position Location: Theory, Practice, and Advances, IEEE, 2019.

[10] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” IEEE Trans. Acoust., Speech, Signal Process., vol. 24, no. 4, pp. 320–327, Aug. 1976.

[11] B. Loesch and B. Yang, “Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions,” in Proc. Int. Conf. Latent Variable Analysis and Signal Separation, Sep. 2010, pp. 41–48.

[12] T. Padois, O. Doutres, and F. Sgard, “On the use of modified phase transform weighting functions for acoustic imaging with the generalized cross correlation,” J. Acoust. Soc. Am., vol. 145, no. 1, pp. 1546–1555, Mar. 2019.

[13] R. Schmidt, “Multiple emitter location and signal parameter estimation,” IEEE Trans. Antennas Propag., vol. 34, no. 3, pp. 276–280, Mar. 1986.

[14] R. Roy and T. Kailath, “ESPRIT-estimation of signal parameters via rotational invariance technique,” IEEE Trans. Acoust., Speech, Signal Process., vol. 37, no. 7, pp. 984–995, Jul. 1989.

[15] R. P. E. D. Di Claudio and G. Orlandi, “Multi-source localization in reverberant environments by ROOF-MUSIC and clustering,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., vol. 2, 2000, pp. 921–924.

[16] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, “Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2011, pp. 117–120.

[17] D. Ying, R. Zhou, J. Li, and Y. Yan, “Window-dominant signal subspace methods for multiple short-term speech source localization,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 25, no. 4, pp. 731–744, 2017.

[18] S. Hafizi, A. H. Moore, and P. A. Naylor, “Multi-source localization using estimation consistency in the time-frequency domain,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2017, pp. 516–520.

[19] L. Wang, T. Hon, J. D. Reiss, and A. Cavallaro, “An iterative approach to source counting and localization using two distant microphones,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 24, no. 6, pp. 1079–1093, Jun. 2016.

[20] B. Yang, H. Liu, C. Pang, and X. Li, “Multiple sound source counting and localization based on TF-Wise spatial spectrum clustering,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 27, no. 8, pp. 1241–1255, 2019.

[21] C. Blandin, A. Ozerv, and E. Vincent, “Multi-source TDOA estimation in reverberant audio using angular spectra and clustering,” Signal Process., vol. 92, no. 8, pp. 1950–1960, 2012.

[22] O. Schwartz, Y. Dorfan, E. Habets, and et al., “Multi-speaker DOA estimation in reverberation conditions using expectation-maximization,” in Proc. IEEE International Workshop on Acoustic Signal Enhancement, 2016, pp. 1–5.

[23] O. Schwartz, Y. Dorfan, M. Taseska, et al., “DOA estimation in noisy environment with unknown noise power using the em algorithm,” in Proc. Hands-free Speech Communications and Microphone Arrays, 2017, pp. 322–326.

[24] A. Brendel, B. Lauer-Goldschtein, S. Gannot, and et al., “Localization of an unknown number of speakers in adverse acoustic conditions using reliability information and diarization,” in Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 7898–7902.

[25] Z.-Q. Wang, X. Zhang, and D. Wang, “Robust TDOA estimation based on time-frequency masking and deep neural networks,” in Proc. Interspeech, 2018, pp. 322–326.

[26] L. Comanducci, M. Cobos, F. Antonacci, and et al., “Time difference of arrival estimation from frequency-sliding generalized cross-correlations using convolutional neural networks,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., Barcelona, Spain, 2020, pp. 4945–4949.

[27] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, “Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates,” Sensors, vol. 18, pp. 10345–10354, Dec. 2018.

[28] P. Pertilla, M. Parviainen, V. Myllylä, and et al., “Time difference of arrival estimation with deep learning - From acoustic simulations to recorded data,” in Proc. IEEE Int. Workshop on Multimedia Signal Process., Tampere, Finland, 2020, pp. 1–6.

[29] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” IEEE Trans. Acoust., Speech, Signal Process., vol. 32, no. 6, pp. 1109–1121, Dec. 1994.

[30] M. Souden, J. Chen, J. Benesty, and S. Affes, “Gaussian model-based multichannel speech presence probability,” IEEE Trans. Audio, Speech, Language Process., vol. 18, no. 5, pp. 1072–1077, Jul. 2010.

[31] T. K. Moon, “The expectation-maximization algorithm,” IEEE Signal Process. Mag., vol. 13, no. 6, pp. 47–60, Nov. 1996.

[32] B. M. Mahmmod and et al., “Speech enhancement algorithm based on super-gaussian modeling and orthogonal polynomials,” IEEE Access, vol. 7, no. 10, pp. 103485–103504, 2019.

[33] H. Krim and M. Viberg, “Two decades of array signal processing research: the parametric method,” IEEE Signal Process. Mag., vol. 13, no. 4, pp. 67–94, 1996.
[34] H. Ye and D. DeGroat, “Maximum likelihood DOA estimation and asymptotic Cramer-Rao bounds for additive unknown colored noise,” *IEEE Trans. Signal Process.*, vol. 43, no. 4, pp. 938–949, 1995.

[35] Y. Izumi, N. Ono, and S. Sagayam, “Sparseness-based 2CH BSS using the EM algorithm in reverberant environment,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 2007, pp. 147–150.

[36] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[37] E. Lehmann and A. Johansson, “Prediction of energy decay in room impulse responses simulated with an image-source model,” *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 269–277, Jul. 2008.

ZHIHUA LU received the B.Sc. degree from Zhejiang University, China, in 2004. He received the Dipl.-Ing. and Dr.-Ing. degrees from Technische Universität Darmstadt, Germany, in 2009 and 2012, respectively. From 2013 to 2015, he was with Temasek Laboratories at Nanyang Technological University in Singapore. Now he is with Ningbo University in China. His research interests include statistical signal processing, sensor array processing, and audio signal processing.

JOÃO P. J. DA COSTA received the Diploma degree in electronic engineering in 2003 from the Military Institute of Engineering (IME), Rio de Janeiro, Brazil, his M.Sc. degree in electrical engineering in 2006 from University of Brasilia (UnB), Brazil, and his Ph.D. degree in electrical and information engineering in 2010 at Ilmenau University of Technology (TU Ilmenau), Germany. From 2013 to 2014, he worked as a Research Associate with the Center for Marine Environmental Sciences at University of Bremen, Bremen, Germany. Since 2014, has been working as a development engineer at HELLA GmbH & Co. KGaA, Lippstadt, Germany, where he is mainly responsible for the development of reliable signal processing algorithms for automotive radar systems.

TAI FEI (S’12-M’19-SM’21) received the B.Eng. degree in telecommunication engineering from Shanghai Maritime University, Shanghai, China, in 2005 and the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering and information technology from Darmstadt University of Technology (TUD), Darmstadt, Germany, in 2009 and 2014, respectively. From 2009 to 2012, he worked as a Research Associate with the Institute of Water-Acoustics, Sonar-Engineering and Signal-Theory at Hochschule Bremen, Bremen, Germany, in collaboration with the Signal Processing Group at TUD, Darmstadt, Germany, where his research interest was the detection and classification of underwater mines in sonar imagery. From 2013 to 2014, he worked as a Research Associate with the Center for Marine Environmental Sciences at University of Bremen, Bremen, Germany. Since 2014, has been working as a development engineer at HELLA GmbH & Co. KGaA, Lippstadt, Germany, where he is mainly responsible for the development of reliable signal processing algorithms for automotive radar systems.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/