INFORMED DIRECTION OF ARRIVAL ESTIMATION USING A SPHERICAL-HEAD MODEL FOR HEARING AID APPLICATIONS

Mojtaba Farmani[1] *Michael Syskind Pedersen*[2] *Zheng-Hua Tan*[1] *Jesper Jensen*[1,2]

[1]Department of Electronic Systems, Aalborg University, {mof, zt, jje}@es.aau.dk
[2]Oticon A/S, Denmark, {msp, jsj}@oticon.dk

ABSTRACT

In this paper, we propose a Direction of Arrival (DoA) estimator for a Hearing Aid System (HAS) which can connect to a wireless microphone worn by a target talker. The wireless microphone "informs" the HAS about the almost noise-free content of the target sound, and the proposed DoA estimator uses the knowledge of the noise-free target sound and the received microphone signals to estimate the DoA via a maximum likelihood approach. Moreover, the proposed DoA estimator resorts to a user-independent spherical-head model to consider the acoustic impacts of the head on the received signals at the HAS. Further, the proposed DoA estimator uses an Inverse Discrete Fourier Transform (IDFT) technique to evaluate the likelihood function computationally efficiently. We assessed the performance of the proposed estimator for various DoAs, Signal to Noise Ratios (SNRs), and target distances in different noisy and reverberant situations. The proposed estimator improves the performance markedly over other recently proposed "informed" DoA estimators.

Index Terms— Sound source localization, Direction of Arrival, Maximum Likelihood, Hearing Aid Systems

1. INTRODUCTION

Direction of Arrival (DoA) estimation of a target sound has been investigated with different approaches in various applications, such as robotics [1–3], video conferencing [4], surveillance [5], wireless acoustic sensor network [6], and hearing aids [7–10]. In this paper, we propose a DoA estimator for an advanced Hearing Aid System (HAS) which can connect to a wireless microphone worn by a target talker. Recognizing the target sound DoA allows HASs to enhance the spatial hearing of the HAS user by maintaining or accentuating the spatial cues of the target sound [10–12].

Most DoA estimation algorithms have been proposed for applications which are "uninformed" about the noise-free content of the target sound, e.g. [1–7, 12–15]. However, advances in wireless technology enable new HASs—where the target talker is wearing a wireless microphone—to have access to an essentially noise-free version of the target signal [8–11]. This change introduces the "informed" DoA estimation problem considered in this paper (Fig. 1).

The "informed" DoA estimation problem was first studied and tackled via a binaural Time-Difference-of-Arrival (TDoA)-based method in [10]. This method estimates the TDoA by resorting to a cross-correlation technique and then maps the estimated TDoA to a DoA estimate through a sine law. This method [10] has a low computational overhead and confines the target locations to the front-horizontal plane.

In previous papers [8,9], we also dealt with the "informed" DoA estimation problem. Specifically, we proposed a maximum likelihood (ML) framework that utilizes the wirelessly transmitted signal and ambient noise characteristics for DoA estimation. The algorithm proposed in [8]—called MLSSL (Maximum Likelihood Sound Source Localization)—uses a database of measured Head Related Transfer functions (HRTFs) of the specific HAS user in order to model the user's head shadowing effect and the acoustic channel. On the other hand, the estimator proposed in [9], which is a TDoA-based DoA estimator, employs a free-field and far-field model to avoid user-related prior assumptions. The signal model in [9] enabled the use of Inverse Discrete Fourier Transform (IDFT) techniques to evaluate the likelihood function computationally efficiently.

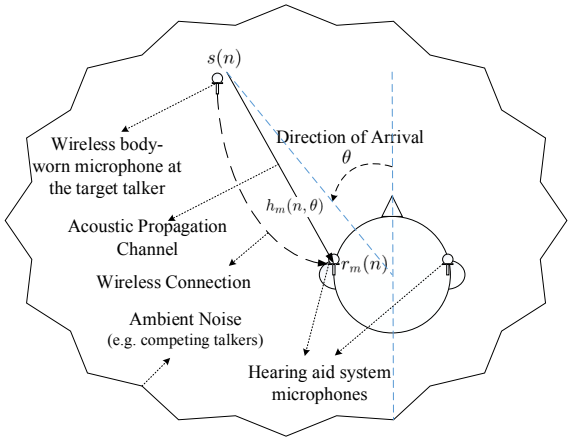


Fig. 1: An "informed" DoA estimation scenario for a hearing aid system using a wireless microphone. $r_m(n)$, $s(n)$ and $h_m(n,\theta)$ are the noisy received sound at microphone m, the noise-free target sound and the acoustic channel impulse response between the target talker and microphone m, respectively. $s(n)$ is available at the hearing aid via wireless connection, and the goal is to estimate θ.

MLSSL [8] and the TDoA-based method [9] form a family of ML-based methods for solving the "informed" DoA estimation problem. These two methods are the two extremes in this family regarding modeling of and dependence on the acoustic characteristics of the specific user's head: MLSSL [8] relies on detailed knowledge of the head characteristics of a specific user, while the TDoA-based method [9] totally ignores the acoustic shadowing effect of the head. In general, MLSSL is more accurate than the TDoA-based method at the cost of higher computation and prior knowledge of HRTFs.

In this paper, we propose an intermediate approach to gain advantages of both methods. To improve the accuracy over the TDoA-based method, we propose a simplified spherical-head model which allows to consider the acoustic effects of the head without being user-dependent. Further, we show that the likelihood function in the proposed method can be computed efficiently using IDFTs. The proposed method is different from [10] because it uses a maximum likelihood approach, which considers the background noise charac-

teristics, models the presence of the head, and estimates the DoA and the TDoA jointly.

2. SIGNAL MODEL

Regarding Fig. 1, for microphone m of the HAS, we can write:

$$r_m(n) = s(n) * h_m(n,\theta) + v_m(n), \qquad m \in \{\text{left, right}\}\ , \quad (1)$$

where r_m, s, h_m and v_m are the noisy signal received at microphone m, the noise-free target signal emitted at the target talker's position, the acoustic channel impulse response between the target talker and microphone m, and an additive noise component, respectively. n is the discrete time index, and $*$ is the convolution operator.

Let $R_m(l,k)$, $S(l,k)$ and $V_m(l,k)$ denote the short time Fourier transform (STFT) of r_m, s and v_m, respectively. Specifically, let

$$R_m(l,k) = \sum_n r_m(n) w(n - lA) \mathrm{e}^{-\frac{j2\pi k}{N}(n-lA)}, \quad (2)$$

where l and k are frame and frequency bin indexes, respectively, N is the frame length, A is the decimation factor, $w(n)$ is the windowing function, and $j = \sqrt{-1}$ is the imaginary unit. We define $S(l,k)$ and $V_m(l,k)$ similarly. Moreover, let $H_m(k,\theta)$ denote the Discrete Fourier Transform (DFT) of h_m:

$$\begin{aligned} H_m(k,\theta) &= \sum_n h_m(n,\theta) \mathrm{e}^{-\frac{j2\pi kn}{N}} && (3)\\ &= \alpha_m(k,\theta) \mathrm{e}^{-\frac{j2\pi k}{N} D_m(k,\theta)}, && (4)\end{aligned}$$

where N is the DFT order, $\alpha_m(k,\theta)$ is a real number and denotes the frequency-dependent attenuation factor due to propagation effects, and $D_m(k,\theta)$ is the frequency-dependent propagation time from the target sound source to microphone m. For simplicity and to decrease computation overhead, we model the acoustic channel as a function that delays and attenuates its input signals uniformly across frequencies [9], i.e.

$$\tilde{H}_m(k,\theta) = \tilde{\alpha}_m(\theta) \mathrm{e}^{-\frac{j2\pi k}{N}\tilde{D}_m(\theta)}, \quad (5)$$

where $\tilde{D}_m(\theta)$ and $\tilde{\alpha}_m(\theta)$ are frequency-independent. Now, we can approximate Eq.(1) in the STFT domain as:

$$R_m(l,k) = S(l,k)\tilde{H}_m(k,\theta) + V_m(l,k). \quad (6)$$

The vector form of Eq. (6) is written as:

$$\boldsymbol{R}(l,k) = S(l,k)\tilde{\boldsymbol{H}}(k,\theta) + \boldsymbol{V}(l,k), \quad (7)$$

where

$$\begin{aligned} \boldsymbol{R}(l,k) &= [\, R_{\text{left}}(l,k),\ R_{\text{right}}(l,k)\,]^{\intercal},\\ \tilde{\boldsymbol{H}}(k,\theta) &= [\, \tilde{H}_{\text{left}}(k,\theta),\ \tilde{H}_{\text{right}}(k,\theta)\,]^{\intercal},\\ \boldsymbol{V}(l,k) &= [\, V_{\text{left}}(l,k),\ V_{\text{right}}(l,k)\,]^{\intercal}, \end{aligned}$$

and the superscript $\intercal$ is the transpose operator.

3. MAXIMUM LIKELIHOOD FRAMEWORK

To define the likelihood function, we assume the additive noise observed at the microphones is distributed according to a zero-mean circularly-symmetric complex Gaussian distribution, i.e. $\boldsymbol{V}(l,k) \sim \mathcal{N}(\boldsymbol{0}, \mathbf{C}_v(l,k))$, where $\mathbf{C}_v(l,k) = \mathrm{E}\{\boldsymbol{V}(l,k)\boldsymbol{V}^{\mathrm{H}}(l,k)\}$, and where $\mathrm{E}\{.\}$ and superscript H represent the expectation and Hermitian transpose operators, respectively. Since $S(l,k)$ is available at the HAS, we can relatively easily determine the time-frequency regions in the received noisy microphone signals, where the target speech is essentially absent; therefore, we adaptively estimate $\mathbf{C}_v(l,k)$ using exponential smoothing over these time-frequency regions. Moreover, we assume the noisy observations are independent across frequencies; therefore, the likelihood function for each frame is defined by:

$$p(\underline{\underline{\mathbf{R}}}(l) | \boldsymbol{S}(l), \underline{\underline{\tilde{\mathbf{H}}}}(\theta), \underline{\underline{\mathbf{C}_v}}(l)) = \prod_{k=1}^{N} \frac{1}{\pi^M |\mathbf{C}_v(l,k)|} e^{\{-(\boldsymbol{Z}(l,k))^{\mathrm{H}} \mathbf{C}_v^{-1}(l,k)(\boldsymbol{Z}(l,k))\}}, \quad (8)$$

where $|.|$ denotes the matrix determinant, N is the number of frequency indexes and

$$\begin{aligned} \underline{\underline{\mathbf{R}}}(l) &= [\, \boldsymbol{R}(l,1),\ \boldsymbol{R}(l,2),\ \cdots,\ \boldsymbol{R}(l,N)\,],\\ \boldsymbol{R}(l,k) &= [\, R_{\text{left}}(l,k),\ R_{\text{right}}(l,k)\,]^{\intercal},\ 1 \le k \le N,\\ \boldsymbol{S}(l) &= [\, S(l,1),\ S(l,2),\ \cdots,\ S(l,N)\,]^{\intercal},\\ \underline{\underline{\tilde{\mathbf{H}}}}(\theta) &= [\, \tilde{\boldsymbol{H}}(1,\theta),\ \tilde{\boldsymbol{H}}(2,\theta),\ \cdots,\ \tilde{\boldsymbol{H}}(N,\theta)\,],\\ \tilde{\boldsymbol{H}}(k,\theta) &= [\, \tilde{H}_{\text{left}}(k,\theta),\ \tilde{H}_{\text{right}}(k,\theta)\,]^{\intercal}\\ &= \begin{bmatrix} \tilde{\alpha}_{\text{left}}(\theta) e^{-j2\pi \frac{k}{N}\tilde{D}_{\text{left}}(\theta)} \\ \tilde{\alpha}_{\text{right}}(\theta) e^{-j2\pi \frac{k}{N}\tilde{D}_{\text{right}}(\theta)} \end{bmatrix},\ 1 \le k \le N,\\ \underline{\underline{\mathbf{C}_v}}(l) &= [\, \mathbf{C}_v(l,1),\ \mathbf{C}_v(l,2),\ \cdots,\ \mathbf{C}_v(l,N)\,]^{\intercal},\\ \boldsymbol{Z}(l,k) &= \boldsymbol{R}(l,k) - S(l,k)\tilde{\boldsymbol{H}}(k). \end{aligned}$$

The corresponding reduced log-likelihood function, with terms independent of θ omitted, is given by:

$$\tilde{\mathcal{L}} = \sum_{k=1}^{N} \{-(\boldsymbol{Z}(l,k))^{\mathrm{H}} \mathbf{C}_v^{-1}(l,k)(\boldsymbol{Z}(l,k))\}. \quad (9)$$

4. DOA ESTIMATION USING A HEAD MODEL

In this section, we aim to find the MLE of θ. The first step is to describe the acoustic model of the head.

4.1. Spherical-head model

To describe the acoustic characteristics of a head, we use the "Inter-Microphone Time Difference" (IMTD) and the "Inter-Microphone Level Difference" (IMLD), which are defined as follows:

$$\begin{aligned} \text{IMTD} : \Delta T(\theta) &= \tilde{D}_{\text{left}}(\theta) - \tilde{D}_{\text{right}}(\theta), && (10)\\ \text{IMLD} : \Delta L(\theta) &= 20\log_{10}\left(\frac{\tilde{\alpha}_{\text{left}}(\theta)}{\tilde{\alpha}_{\text{right}}(\theta)}\right), && (11)\end{aligned}$$

where $\tilde{D}_m$ and $\tilde{\alpha}_m$ are defined in Eq. (5).

In general, IMTD and IMLD are frequency-dependent; however, to compute the likelihood function computationally efficiently using IDFTs, we assume they are frequency-independent. Despite this crude assumption, we show in our simulation experiments that this leads to performance improvements. For a rigid spherical head, the IMTD can be approximated by [16]:

$$\text{IMTD} : \tilde{D}_{\text{left}}(\theta) - \tilde{D}_{\text{right}}(\theta) = \frac{b}{c}\left(\sin(\theta) + \theta\right), \quad (12)$$

where b is the sphere radius and c is the speed of sound. To model the IMLD, we use the following relation inspired by the work in [15]:

$$\text{IMLD} : 20\log_{10}\left(\frac{\tilde{\alpha}_{\text{left}}(\theta)}{\tilde{\alpha}_{\text{right}}(\theta)}\right) = \gamma \sin(\theta). \quad (13)$$

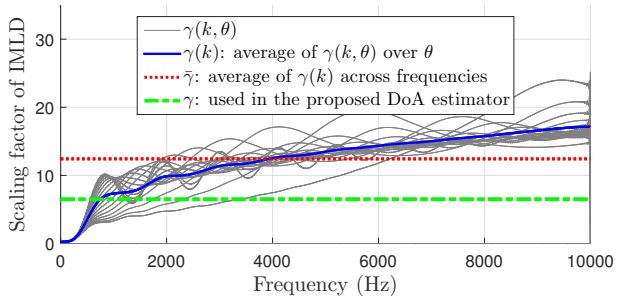


Fig. 2: Scaling factor γ of IMLD (Eq. (13)) for a spherical head using theoretical HRTFs [17].

In [15], γ is a frequency-dependent scaling factor, which is generally smaller at lower frequencies and larger at higher frequencies; however, to be able to apply IDFTs, we assume γ to be frequency-independent. We describe how to determine this value in sec. 4.3.

4.2. DoA estimator

To find the MLE of θ, we expand Eq. (9). Let us denote

$$\mathbf{C}_v^{-1}(l,k) \equiv \begin{bmatrix} C_{11}(l,k) & C_{12}(l,k) \\ C_{21}(l,k) & C_{22}(l,k) \end{bmatrix}. \tag{14}$$

From Eqs. (12) and (13), $\tilde{D}_{\text{right}}$ and $\tilde{\alpha}_{\text{right}}$ can be expressed in terms of $\tilde{D}_{\text{left}}$ and $\tilde{\alpha}_{\text{left}}$, respectively. Inserting these expressions in Eq. (9), we arrive at $\tilde{\mathcal{L}}(\theta, \tilde{D}_{\text{left}}, \tilde{\alpha}_{\text{left}})$ which is independent of $\tilde{D}_{\text{right}}$ and $\tilde{\alpha}_{\text{right}}$. To eliminate the dependency on $\tilde{\alpha}_{\text{left}}$, we insert the MLE of $\tilde{\alpha}_{\text{left}}$ in $\tilde{\mathcal{L}}$. It can be shown that the MLE of $\tilde{\alpha}_{\text{left}}$ is:

$$\hat{\alpha}_{\text{left}} = \frac{f(\theta, D_{\text{left}})}{g(\theta)}, \tag{15}$$

where

$$\begin{aligned} f(\theta, D_{\text{left}}) = \sum_{k=1}^{N} \Bigg(& C_{11}(l,k) R_{\text{left}}(l,k) + \\ & C_{12}(l,k) R_{\text{right}}(l,k) + 10^{\frac{\gamma \sin(\theta)}{20}} \Big(C_{21}(l,k) R_{\text{left}}(l,k) + \\ & C_{22}(l,k) R_{\text{right}}(l,k) \Big) e^{j2\pi \frac{k}{N} [-\frac{b}{c}(\sin(\theta)+\theta)]} \Bigg) \times \\ & S^*(l,k) e^{j2\pi \frac{k}{N} D_{\text{left}}(\theta)}, \end{aligned} \tag{16}$$

$$\begin{aligned} g(\theta) = \sum_{k=1}^{N} \Big(& C_{11}(l,k) + \\ & 2 \times 10^{\frac{\gamma \sin(\theta)}{20}} C_{21} e^{j2\pi \frac{k}{N} [-\frac{b}{c}(sin(\theta)+\theta)]} + \\ & 10^{\frac{\gamma \sin(\theta)}{10}} C_{22}(l,k) \Big) |S(l,k)|^2, \end{aligned} \tag{17}$$

where [.] rounds to nearest integer. Inserting $\hat{\alpha}_{\text{left}}$ into $\tilde{\mathcal{L}}$ gives us:

$$\tilde{\mathcal{L}}(\theta, D_{\text{left}}) = \frac{f^2(\theta, D_{\text{left}})}{g(\theta)}. \tag{18}$$

Note that $f(\theta, D_{\text{left}})$ in Eq. (16) has a structure of an IDFT, which can be evaluated computationally efficiently, with respect to D_{left}; therefore, for a given θ, computing $\tilde{\mathcal{L}}(\theta, D_{\text{left}})$ results in a discrete-time sequence, where the MLE of D_{left} is the time index of the maximum of the sequence. Since θ is unknown, we consider a discrete set Θ of different θs, and evaluate $\tilde{\mathcal{L}}(\theta, D_{\text{left}})$ using an IDFT for each $\theta \in \Theta$. The MLEs of D_{left} and θ are then given by the global maximum:

$$[\hat{\theta}, \hat{D}_{\text{left}}] = \arg\max_{\theta \in \Theta, D_{\text{left}}} \tilde{\mathcal{L}}(\theta, D_{\text{left}}). \tag{19}$$

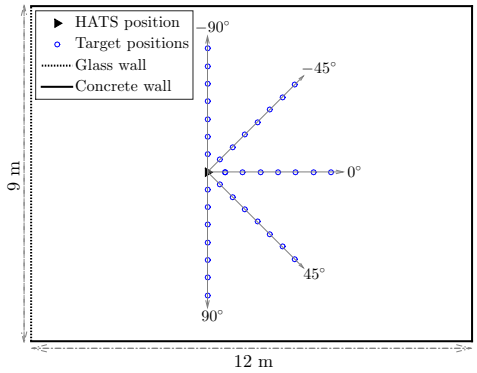


Fig. 3: The map of the room used for HRIRs measurements.

4.3. Scaling factor γ

The only remaining issue is the value of γ, which should be inserted in Eqs. (16) and (17) to evaluate Eq. (18). As shown in Fig. 2, ideally, the scaling factor is frequency- and DoA-dependent ($\gamma(k,\theta)$). To find a frequency- and DoA-independent γ, one could consider averaging over DoAs and frequencies, which leads to $\bar{\gamma} \approx 12.4$. However, in the considered application, the target signal is speech, which is a relatively low-pass signal. Therefore, we expect that low-frequency components should play a larger role in finding γ.

To find the appropriate value of γ, we run simulations for numerous acoustic setups and different $\gamma \in \Gamma = \{1, 1.5, 2, ..., 20\}$ and select the γ leading to the best DoA estimation performance. We evaluate the performance in terms of Mean Absolute Error (MAE):

$$\sigma = \frac{1}{\mathrm{L}} \sum_{j=1}^{\mathrm{L}} |\theta - \hat{\theta}_j|, \tag{20}$$

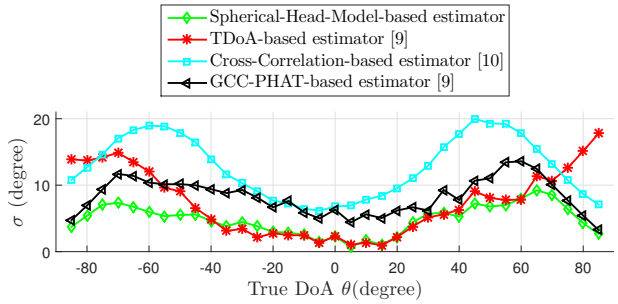
where $\hat{\theta}_j$ is the estimated DoA for the j^{th} frame of the signal, and L is the number of target-active frames. We use the value of γ which minimizes the MAE over the considered conditions.

To simulate a rigid spherical-head, we use theoretical HRTFs proposed in [17]. We run simulations for 72 different configurations: four different target sources (two males and two females), three different distances (1 m, 5 m and 10 m), three different SNRs (-10 dB, 0 dB and 10 dB) and two different noise types (large-crowd noise and bottling-factory-hall noise). The signal duration for each configuration is 60 s, and we use the speech database provided by [18] for the target signals. For each configuration, the target source is placed at 35 different angles at the front-horizontal plane, i.e. $\theta \in \{-85°, -80°, \cdots, 85°\}$. The other simulation parameters are as follows: the sampling frequency is 20 kHz, $N = 2048$, $A = 1024$, and $w(n)$ is a Hamming window.

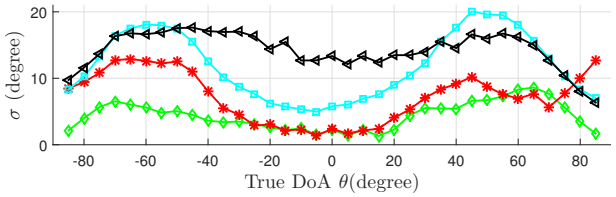
From the simulation results, we find that $\gamma = 6.5$ provides minimum MAE averaged over all considered configurations and θs. As expected, the obtained value of $\gamma = 6.5$ is less than the result of a simple averaging of the scaling factor over the frequencies for the considered spherical head, i.e. $\bar{\gamma} \approx 12.4$ (Fig. 2).

5. SIMULATION RESULTS

In this section, we evaluate the proposed estimator under realistic conditions which were not used in the simulation experiments to find γ. Here, we study the impacts of the true DoA, noise type, SNR, reverberation level, and the target distance on the performance of the proposed estimator. In the following, the proposed estimator is referred as "Spherical-Head-Model-based DoA estimator".



(a) Babble-speech noise.



(b) Bottling-factory-hall noise.

Fig. 4: Performance as a function of θ at SNR = 0 dB in an anechoic room.

5.1. Setup

To simulate real world scenarios, we use two different sets of Head Related Impulse Responses (HRIRs) measured with behind-the-ear (BTE) hearing aids mounted behind each pinna of a head-and-torso-simulator (HATS). The first set of HRIRs was measured in an anechoic chamber for 35 positions uniformly spaced on a semicircle in the front-horizontal plane with radius 1 m centered at the HATS, i.e. $\theta \in \{-85^\circ, -80^\circ, ..., 85^\circ\}$. The second set was measured in a reverberant room shown in Fig. 3. These HRIRs were measured for 35 positions: five DoAs $\theta \in \{-90^\circ, -45^\circ, 0^\circ, 45^\circ, 90^\circ\}$ versus seven distances $d \in \{0.5\,\text{m}, 1\,\text{m}, 1.5\,\text{m}, ..., 3.5\,\text{m}\}$. To simulate a signal from a position, the signal is convolved with its related HRIR.

As target signal, we consider a four-minute signal composed of two male and two female speech signals [18]. We consider two different noise-types: speech-babble and bottling-factory-hall noise. Speech-babble is synthesized by playing back different speech signals from each θ simultaneously. The TSP database [18], which consists of different male and female voices, is used as noise sources. The wide-band SNR in each simulation experiments is expressed relative to the left-ear microphone signals. The other simulation parameters are as follows: the sampling frequency is 20 kHz, $N = 2048$, $A = 1024$, $w(n)$ is a Hamming window, the length of $w(n)$ and the DFT order are the same, and $\Theta = \{-90^\circ, -85^\circ, \cdots, 90^\circ\}$. We use the MAE (Eq. (20)) as performance metric.

5.2. Results and discussion

Fig. 4 shows the MAE of various "informed" DoA estimators as a function of θ at an SNR of 0 dB for two different noise-types in an anechoic room. Clearly, the proposed spherical-head-model-based estimator performs better than existing "informed" DoA estimators, and appears robust against the noise types. In contrast, the performance of the "informed" GCC-PHAT-based estimator, introduced in [9], is quite dependent on the noise types. As mentioned before, the TDoA-based estimator [9] relies on a free-field assumption, which is more valid for $\theta \approx 0^\circ$ and less valid for $\theta \approx \pm 90^\circ$. The influence of the free-field assumption is clearly visible in the results of the TDoA-based estimator. On the other hand, because the proposed estimator simulates the presence of the head, it improves the performance of the DoA estimation compared with the TDoA-based estimator for $\theta \in [-90, -50]$ or $\theta \in [60, 90]$.

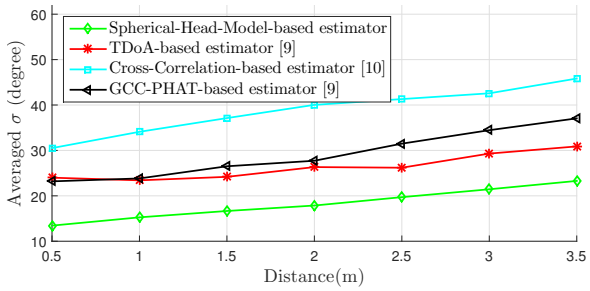


Fig. 5: Performance as a function of distance in a reverberant room shown in Fig. 3 at SNR = 0 dB, and σ is averaged over all θs.

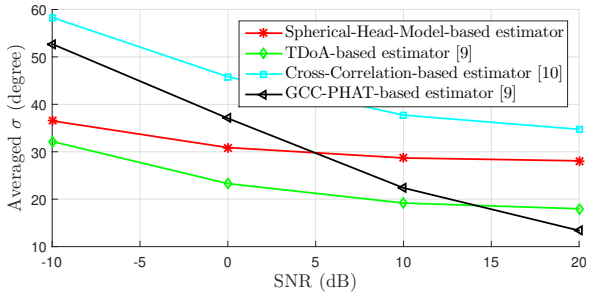


Fig. 6: Performance as a function of SNR in a reverberant room shown in Fig. 3. $d = 3.5$m, and σ is averaged over all θs.

Fig. 5 shows the MAE of the estimators averaged across the noise types and θs as a function of target distance in a reverberant room (Fig. 3). In general, increasing the distance will decrease the direct-to-reverberant energy ratio [19], i.e. reverberation will degrade the received signals more at larger distances. However, the proposed estimator still shows consistent improvement.

Fig. 6 shows the MAE of the estimators averaged across the noise types and θs as a function of SNR in a reverberant situation. As expected, the higher the SNR, the better the performance. The excellent performance of the GCC-PHAT-based estimator at high SNRs may be explained by the fact that the PHAT algorithm is almost ML optimal in low-noise reverberant environments [20]. While the proposed method already performs decently in this situation, we expect that a signal model which directly takes the reverberation into account, e.g. [4, 21], would improve the performance further.

6. CONCLUSION

In this paper, we proposed a DoA estimator for a hearing aid system which has access to the noise-free target signal via a wireless microphone. We employed a spherical-head model and proposed a maximum likelihood approach to estimate the DoA. We showed that the considered signal model allowed the likelihood function to be calculated efficiently via Inverse-Discrete- Fourier-Transform techniques. In simulation experiments, we studied the effects of the true DoA, noise type, SNR, reverberation level, and target distance on the performance of the proposed algorithm. The proposed method improves the estimation performance over recently proposed "informed" DoA estimators, especially, when the target is at the sides of the head, where the influence of a head model is largest.

7. REFERENCES

[1] J. A. Macdonald, "A localization algorithm based on head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4290–4296, June 2008.

[2] C. Vina, S. Argentieri, and M. Rébillat, "A spherical cross-channel algorithm for binaural sound localization," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 2921–2926.

[3] F. Keyrouz, "Advanced binaural sound localization in 3-D for humanoid robots," *IEEE Transaction on Instrumentation and Measurement*, vol. 63, no. 9, pp. 2098–2107, Sept 2014.

[4] C. Zhang, D. Florencio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[5] J. Kotus, K. Lopatka, and A. Czyzewski, "Detection and localization of selected acoustic events in acoustic field for smart surveillance applications," *Multimedia Tools and Applications*, vol. 68, no. 1, pp. 5–21, 2014.

[6] A. Hassani, A. Bertrand, and M. Moonen, "Cooperative integrated noise reduction and node-specific direction-of-arrival estimation in a fully connected wireless acoustic sensor network," *Signal Processing*, vol. 107, pp. 68–81, Feb. 2015.

[7] S. Goetze, T. Rohdenburg, V. Hohmann, B. Kollmeier, and K.-D. Kammeyer, "Direction of arrival estimation based on the dual delay line approach for binaural hearing aid microphone arrays," in *International Symposium on Intelligent Signal Processing and Communication Systems*, Nov 2007, pp. 84–87.

[8] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Maximum likelihood approach to informed sound source localization for hearing aid applications," in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 439–443.

[9] M. Farmani, M. S. Pedersen, Z. H. Tan, and J. Jensen, "Informed TDoA-based Direction of Arrival estimation for hearing aid applications," in *IEEE Global Conference on Signal and Information Processing*, 2015.

[10] G. Courtois, P. Marmaroli, M. Lindberg, Y. Oesch, and W. Balande, "Implementation of a binaural localization algorithm in hearing aids: specifications and achievable solutions," in *Audio Engineering Society Convention 136*, April 2014, p. 9034.

[11] G. Courtois, P. Marmaroli, H. Lissek, Y. Oesch, and W. Balande, "Binaural hearing aids with wireless microphone systems including speaker localization and spatialization," in *Audio Engineering Society Convention 138*, May 2015, p. 9242.

[12] W. Wu, C. Hsieh, H. Huang, and O. T.-C. Chen, "Hearing aid system with 3D sound localization," in *IEEE Region 10 Conference TENCON*, Oct. 2007, pp. 1–4.

[13] A. Pourmohammad and S. M. Ahadi, "Real time high accuracy 3-D PHAT-based sound source localization using a simple 4-microphone arrangement," *IEEE Systems Journal*, vol. 6, no. 3, pp. 455–468, Sept. 2012.

[14] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer Speech & Language*, vol. 11, no. 2, pp. 91–126, 1997.

[15] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, 2010.

[16] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a spherical-head model from anthropometry," *Journal of Audio Engineering Society*, vol. 49, no. 6, pp. 472–479, 2001.

[17] R. Duda and W.L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, 1998.

[18] P. Kabal, "TSP speech database," Tech. Rep., Department of Electrical and Computer Engineering, McGill University, 2002.

[19] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using d/r spatial correlation matrix model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2374–2384, Nov 2011.

[20] C. Zhang, D. Florêncio, and Z. Zhang, "Why does PHAT work well in low noise, reverberative environments?," in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 2565–2568.

[21] A. Kuklasinski, S. Doclo, T. Gerkmann, S. Holdt Jensen, and J. Jensen, "Multi-channel PSD estimators for speech dereverberation - a theoretical and experimental comparison," in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2015, pp. 91–95.