MAXIMUM LIKELIHOOD APPROACH TO "INFORMED" SOUND SOURCE LOCALIZATION 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

Most state-of-the-art Sound Source Localization (SSL) algorithms have been proposed for applications which are "uninformed" about the target sound content; however, utilizing a wireless microphone worn by a target talker, enables recent Hearing Aid Systems (HASs) to access to an almost noise-free sound signal of the target talker at the HAS via the wireless connection. Therefore, in this paper, we propose a maximum likelihood (ML) approach, which we call MLSSL, to estimate the Direction of Arrival (DoA) of the target signal given access to the target signal content. Compared with other "informed" SSL algorithms which use binaural microphones for localization, MLSSL performs better using signals of one or more microphones placed on just one ear, thereby reducing the wireless transmission overhead of binaural hearing aids. More specifically, when the target location confined to the front-horizontal plane, MLSSL shows an average absolute DoA estimation error of 5 degrees at SNR of -5 dB in a large-crowd noise and non-reverberant situation. Moreover, MLSSL suffers less from front-back confusions compared with the recent approaches.

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

1. INTRODUCTION

Sound Source Localization (SSL) has been investigated in many applications, such as robotics [1, 2, 3], video conferencing [4], and hearing aids [5]. In a sense, SSL is a primitive task which would improve performance of higher level tasks. For example, in a Hearing Aid System (HAS), knowing the location of the target sound may improve noise reduction algorithms [6, 7], leading to better speech enhancement performance.

In general, different acoustic localization strategies using microphone arrays have been investigated [8, ch. 8]:

- Steered-Beamformer-Based Location Estimators: these methods steer the beam to the potential sound source locations and search for a maximum in output power (termed focalization) [9].
- High-Resolution-Spectral-Estimation-Based Location Estimators: these methods exploit the spatiospectral correlation matrix obtained from the microphones signals. Under certain assumptions, the sound source locations can be derived from a lower-dimensional vector subspace embedded within the signal space spanned by the columns of the correlation matrix [8, ch. 8].
- Time-Difference-of-Arrival (TDoA)-Based Location Estimators: these methods use a set of TDoA estimations of the signals reaching each pair of microphones to estimate the sound source location [8, ch. 8][10].

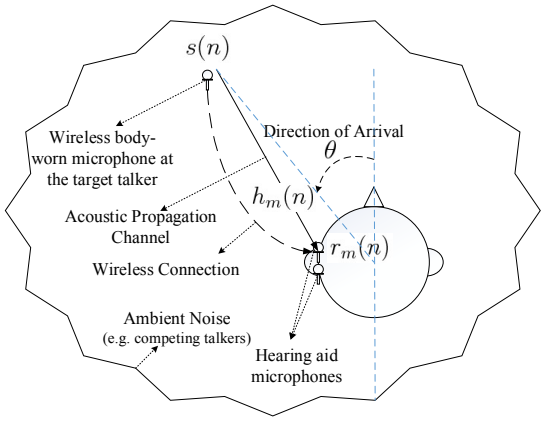


Fig. 1: SSL scenario for a hearing aid system using a wireless microphone: $r_m(n)$, $s(n)$ and $h_m(n)$ are the noisy received sound, the noise-free target sound and the corresponding HRIR for microphone m, respectively. $s(n)$ is available at the hearing aid via wireless connection to the wireless microphone at the target talker. Estimating the direction of arrival θ is the goal in this scenario.

When the microphone array is located next to the ears, like in HASs or humanoid robots, bio-inspired binaural cues, such as Interaural Time Difference (ITD), Interaural Intensity Difference (IID) and monaural cues represented by Head Related Transfer Functions (HRTFs) [called Head Related Impulse Responses (HRIRs) in the time domain] are often used for SSL [11]. Roughly, humans are thought to use ITDs for low frequency components, up to approximately 1500 Hz, and IIDs for higher frequency components [12]. For monaural spatial hearing, humans are believed to utilize the spectral filtering of the incoming sound at the head, torso and pinnae [11], i.e. filtering of the incoming sound through HRTFs.

Most current SSL algorithms have been proposed for applications which are "uninformed" about the target source signal content [1, 3, 4], i.e. they do not have any access to the noise-free target signal content. However, recent advances in wireless technology enables new HASs, where the target talker is wearing a wireless microphone, to have access to an essentially noise-free version of the target signal [5]. This turns the "uninformed" SSL problem into the "informed" SSL problem considered in this paper.

Fig. 1 depicts the system considered in this paper. The target signal $s(n)$ is transmitted through the acoustic channel $h_m(n)$ and reaches the m^{th} microphone of the HAS. Due to additive environmental noise, a noisy signal $r_m(n)$ is received at the m^{th} microphone. Moreover, the noise-free target signal $s(n)$ is also transmitted to the HAS via the wireless connection. We aim at estimating the target signal Direction of Arrival (DoA) θ based on these signals.

In HASs, since microphones are located at the ears, the acoustic shadowing effects of the user's head and torso cause $h_m(m)$ to depend on θ [11]. However, for simplicity, many SSL algorithms, e.g. [5, 10], assume a free field situation and disregard the user's head and torso acoustic shadowing effect, causing the location estimation performance to be reduced. In this paper, we propose a method which does take the head presence into account to distinguish directions, thereby improving localization performance. The proposed method is a maximum likelihood approach; therefore, we call it Maximum Likelihood Sound Source Localization (MLSSL).

2. SIGNAL MODEL

Fig. 1 shows the situation at hand: the noisy received sound signal $r_m(n)$ at microphone m is a result of the convolution of the target signal $s(n)$ with the acoustic channel impulse response $h_m(n)$ from the target talker to microphone m, and is contaminated by additive noise $v_m(n)$. For each microphone of the HAS, we can write:

$$\begin{aligned} r_m(n) &= d_m(n) + v_m(n), \qquad m = 1, \cdots, M\ , & (1)\\ d_m(n) &= s(n) * h_m(n), & (2)\end{aligned}$$

where $M \geq 1$ is the number of available microphones, n is the discrete time index, and $*$ is the convolution operator.

Most state-of-the-art HASs operate in the short time Fourier transform (STFT) domain because it allows frequency dependent processing, computational efficiency and low latency algorithm implementation. Therefore, let

$$\begin{aligned} S(l,k) &= \sum_n s(n)w(n-lA)\mathrm{e}^{-\frac{j2\pi k}{N}(n-lA)}, & (3)\\ D_m(l,k) &= \sum_n \sum_t h_m(t)s(n-t) \times \\ &\qquad w(n-lA)\mathrm{e}^{-\frac{j2\pi k}{N}(n-lA)} \\ &= \sum_n s(n) \sum_t h_m(t) \times \\ &\qquad w(n+t-lA)\mathrm{e}^{-\frac{j2\pi k}{N}(n+t-lA)} & (4)\end{aligned}$$

denote the STFT representations of $s(n)$ and $d_m(n)$, respectively, where l and k are frame and frequency bin indices, 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. Moreover, let

$$H_m(k) = \sum_t h_m(t)\mathrm{e}^{-\frac{j2\pi kt}{N}} \qquad (5)$$

denote the discrete Fourier transform of $h_m(n)$, where N is greater or equal to the duration of $h_m(n)$. Eq. (4) implies that $D_m(l,k) \neq S(l,k)H_m(k)$. However, if the support of $w(n)$ is smoothly long enough compared with the duration of $h_m(n)$, then $w(n-t)h_m(t) \approx w(n)h_m(t)$ [13]; in this case, we find:

$$\begin{aligned} D_m(l,k) &\approx \sum_n s(n)w(n-lA)\mathrm{e}^{-\frac{j2\pi k}{N}(n-lA)} \times \\ &\qquad \sum_t h_m(t)\mathrm{e}^{-\frac{j2\pi k}{N}(t)} & (6)\\ &= S(l,k)H_m(k), & (7)\end{aligned}$$

i.e. $D_m(l,k)$ can be approximated as a point-wise multiplication of $S(l,k)$ and $H_m(l,k)$ [13]. With this approximation, Eq. (1) can be approximated in the STFT domain as:

$$R_m(l,k) = S(l,k)H_m(k) + V_m(l,k), \qquad (8)$$

where $R_m(l,k)$ and $V_m(l,k)$ are STFT coefficients of the received signal and noise signal for the m^{th} microphone, respectively, and are defined analogously to $S(l,k)$ in Eq. (3).

Collecting the M microphone equations (Eq. (8)) in a column vector gives rise to the following signal model:

$$\boldsymbol{R}(l,k) = S(l,k)\boldsymbol{H}(k) + \boldsymbol{V}(l,k), \qquad (9)$$

where

$$\begin{aligned} \boldsymbol{R}(l,k) &= [R_1(l,k), R_2(l,k), \cdots, R_M(l,k)]^{\mathrm{T}}, & (10)\\ \boldsymbol{H}(k) &= [H_1(k), H_2(k), \cdots, H_M(k)]^{\mathrm{T}}, & (11)\\ \boldsymbol{V}(l,k) &= [V_1(l,k), V_2(l,k), \cdots, V_M(l,k)]^{\mathrm{T}}. & (12)\end{aligned}$$

3. MAXIMUM LIKELIHOOD ESTIMATION OF DOA

The acoustic shadowing effects of the head and torso cause $\boldsymbol{H}(k)$ to depend on θ [11]; therefore, if we possess a prestored database $\mathcal{H} = \{\boldsymbol{H}_1, \boldsymbol{H}_2, \cdots, \boldsymbol{H}_I\}$, which consists of I sets of HRTFs labelled by their corresponding θ, the target θ may be estimated by finding the best candidate in $\mathcal{H}$. In fact, $\mathcal{H}$ is a discrete model of the continuous space of HRTFs. To find the best $\boldsymbol{H}_i$ in $\mathcal{H}$ based on the received signals, we introduce a maximum likelihood strategy.

Let us assume that $\boldsymbol{V}(l,k)$ in Eq. (9) is a zero-mean, circularly-symmetric complex Gaussian random vector, i.e. $\boldsymbol{V}(l,k) \sim \mathcal{N}(\boldsymbol{0}, \mathbf{C}_V(l,k))$, where $\mathbf{C}_V(l,k)$ is the inter-microphone noise covariance matrix. Since we assume the target signal is picked up without any noise by the wireless microphone, $S(l,k)$ is available at the HAS, and we consider it as deterministic and known. $\boldsymbol{H}(k)$ is also considered deterministic but unknown ($\boldsymbol{H} \in \mathcal{H}$). Hence, from Eq. (9) follows:

$$\boldsymbol{R}(l,k) \sim \mathcal{N}(S(l,k)\boldsymbol{H}(k), \mathbf{C}_V(l,k)). \qquad (13)$$

Since $S(l,k)$ is available at the HAS, we can relatively easily determine the time-frequency regions in the noisy microphones signals where the target speech is essentially absent; therefore, we adaptively estimate $\mathbf{C}_V(l,k)$ using exponential smoothing over the frames where the noise is dominant. Furthermore, for mathematical convenience, we assume that the noisy observations are independent over time and frequency. Therefore, the likelihood function of each $\boldsymbol{H}_i \in \mathcal{H}$ regarding the received signals at frame l is defined as:

$$f_l(\boldsymbol{R}, S; \boldsymbol{H}_i) = \prod_{j=l-\mathrm{D}+1}^{l} \prod_{k=1}^{\mathrm{K}} \frac{1}{\pi^M |\mathbf{C}_V(j,k)|} \mathrm{e}^{\{-\boldsymbol{Z}_i^{\mathrm{H}}(j,k)\mathbf{C}_V^{-1}(j,k)\boldsymbol{Z}_i(j,k)\}}, \qquad (14)$$

where $\boldsymbol{Z}_i(j,k) = \boldsymbol{R}(j,k) - S(j,k)\boldsymbol{H}_i(k)$, and $|.|$ and $^{\mathrm{H}}$ denotes the matrix determinant and Hermitian transpose operator, respectively. D is the number of frames and K is the number of frequency indices used to compute the likelihood. It should be noted that we assume that the target source location is fixed across D frames. The corresponding log-likelihood function is given by:

$$\begin{aligned} \mathcal{L}_l(\boldsymbol{H}_i) &= -M\mathrm{DK}\log\pi - \sum_{j=l-\mathrm{D}+1}^{l} \sum_{k=1}^{\mathrm{K}} \log|\mathbf{C}_V(j,k)| - \\ &\qquad \sum_{j=l-\mathrm{D}+1}^{l} \sum_{k=1}^{\mathrm{K}} \boldsymbol{Z}_i^{\mathrm{H}}(j,k)\mathbf{C}_V^{-1}(j,k)\boldsymbol{Z}_i(j,k), & (15)\end{aligned}$$

leading to the maximum likelihood estimation of the HRTF:

$$\boldsymbol{H}_{\mathrm{ML}} = \arg\max_{\boldsymbol{H}_i \in \mathcal{H}} \mathcal{L}_l(\boldsymbol{H}_i), \qquad (16)$$

from which the corresponding DoA estimate $\hat{\theta}$ follows. We solve Eq. (16) via an exhaustive search in $\mathcal{H}$.

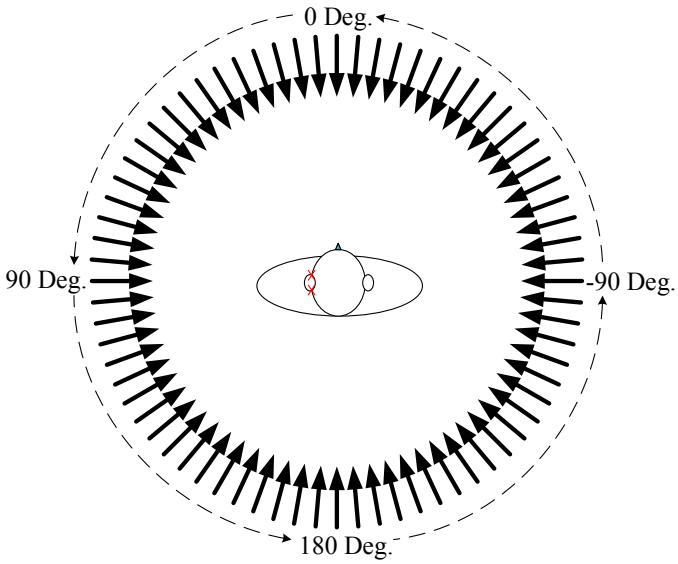


Fig. 2: Experiment setup. In an anechoic chamber, 72 loudspeakers, represented by arrows, are placed on a circle with radius 1.5 m in the horizontal plane centered at the HATS. Microphones locations are represented by $\times$ behind the left ear of the HATS (i.e. around 90°).

4. SIMULATIONS RESULTS

4.1. Experiment setup

Fig. 2 shows the situation considered for assessing the algorithm. The target source is assumed to be placed at one of 72 uniformly spaced possible positions, i.e. with a 5 degrees resolution, on a circle in the horizontal plane with radius 1.5 m centered at a head-and-torso-simulator (HATS). Behind the left pinna of the HATS a two-microphone behind-the-ear (BTE) hearing aid is placed. The distance between front and rear microphones is 12 mm, and the sampling frequency of the microphone signals is 20 kHz. The other simulation parameters are as follows: $N = 2048$ samples, $A = 1024$ samples, and $\mathrm{D} = 2$. $\mathcal{H}$ consists of $I = 72$ sets of HRTFs, measured from each loudspeaker to microphones, and the target speech signal is a 10-second sample of the ISTS signal [14] composed of 21 female voices in 6 different languages. To approximate a practical large-crowd noise field, we play back different speech signals from each of the $I = 72$ target positions simultaneously. The database provided by [15], which consists of different male and female voices, is used as noise sound sources.

When the power of the noise sources is fixed, then the signal-to-noise-ratio (SNR) observed at each of the microphones is a function of θ since the target signal is filtered by the head and torso of the HAS user. Specifically, the SNR is generally reduced when the microphone is in the "shadow" part of the head compared with the case where the microphone is in the "sunny" part. Moreover, for the same θ, "sunny" part microphones have higher SNRs than "shadow" part microphones; therefore, the reference SNRs of the simulations are expressed relative to the Left-Front microphone and $\theta = 0^\circ$.

As performance metrics, we define the percentage of the DoA correct detection and the DoA estimation mean absolute error (MAE) in the following. Let Q_θ denote the number of frames for which $\hat{\theta} = \theta$. The percentage of the DoA correct detections is:

$$P_\theta = \frac{Q_\theta}{\mathrm{L}} \times 100, \tag{17}$$

where L is the total number of frames of the received signals. Moreover, the mean absolute error (MAE) of the DoA estimation is given by:

$$\sigma_{\hat{\theta}} = \frac{1}{\mathrm{L}} \sum_{j=1}^{\mathrm{L}} |\theta - \hat{\theta}_j|, \tag{18}$$

where $\hat{\theta}_j$ is the estimated DoA for the j^{th} frame of the signal.

4.2. MLSSL using one microphone

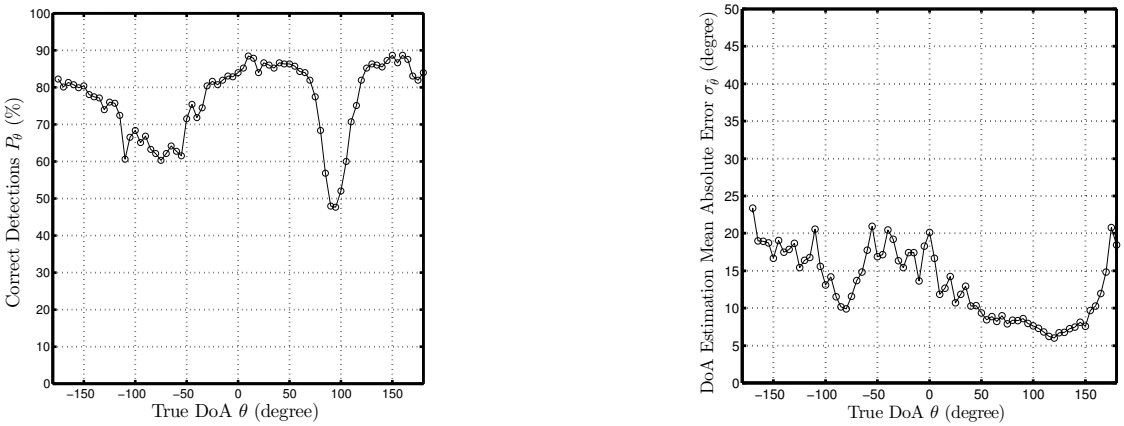
In contrast to other SSL algorithms which often use two microphones, MLSSL allows us to estimate θ with just one microphone. Figs. 3a and 3b show the MLSSL performance in terms of P_θ and $\sigma_{\hat{\theta}}$ at a reference SNR of 0 dB for the full-band signal using $M = 1$ microphone signal (the Left-Front microphone). As can be seen, P_θ drops when the target is located at the sides of the HATS (i.e. $\theta \approx -90^\circ$ and $\theta \approx 90^\circ$), compared with when the target is in front ($\theta \approx 0^\circ$) or behind ($\theta \approx 180^\circ$). On the other hand, $\sigma_{\hat{\theta}}$ in Fig. 3b shows that even though MLSSL has lower P_θ for θ close to -90° or 90°, the MAE is less than the cases where θ is close to 0° or 180°. To explain these behaviours, we plot the MLSSL confusion matrix shown in Fig. 4. Each column of the matrix relates to a θ, and represents the normalized histogram of $\hat{\theta}$s for that particular θ. The almost red diagonal of the matrix shows that MLSSL is generally successful in estimating the θ. However, the two parallel anti-diagonal lines show that when MLSSL fails in detecting the correct θ, then the most probable cause of errors is a front-back confusion. Front-back confusions result in larger estimation errors for the θs in the front or back of the HATS than the left or right sides θs and explain the higher $\sigma_{\hat{\theta}}$ around $\theta = 0^\circ$ or 180°. As mentioned before, the SNR is a function θ and is almost higher for $\theta \approx 90^\circ$ when the microphones are on the left ear, but since influences of the head and torso are small for $\theta \approx 90^\circ$, their HRTFs are locally very similar and cause local errors and relatively low P_θ. Finally, as can be seen in Fig. 3b, $\sigma_{\hat{\theta}}$ is generally higher when $\theta \in [-180^\circ, 0^\circ]$ since the microphone is in the head shadow region, and the SNR is lower.

4.3. Comparison with the state-of-the-art

Courtois et al. in [5] recently introduced the informed SSL problem and proposed a solution based on ITD and binaural signals. They use the wirelessly received noise-free target signal as a time reference to estimate the ITD, and then to estimate θ, they resort to a "sine law" [5]. This causes their method to be unable to differentiate between front and back angles, e.g. $\theta = 45^\circ$ and $\theta = 135^\circ$. Although MLSSL does not have this limitation, for comparison, we consider the frontal horizontal plane only.

Fig. 5 shows $\sigma_{\hat{\theta}}$ for MLSSL using one or two microphones placed behind the left ear, compared with the Courtois et al. method [5]. As can be seen, MLSSL performs significantly better for all θs. The Courtois et al. results are symmetric with respect to $\theta = 0$ since they use binaural signals, but MLSSL results are asymmetric because microphones are located at one ear only, and the head shadow influences signals which are coming from left and right differently. Furthermore, comparing Figs. 5 and 3b shows that knowing a priori that θ is in the frontal plane, improves the MLSSL $\sigma_{\hat{\theta}}$ significantly by eliminating front-back confusions.

In practice, since θ is a continuous variable, it may be represented exactly by none of the HRTFs in $\mathcal{H}$. To assess MLSSL performance in this situation, we made a reduced database $\mathcal{H}'$ by eliminating every other HRTF from $\mathcal{H}$, i.e. there is no HRTF in $\mathcal{H}'$ for half of the considered θs. Fig. 6 shows MAE of the methods averaged over all the frontal θs as a function of SNR. As expected, MLSSL has the best performance when θ is represented in the database. But when θ is not in $\mathcal{H}'$, MLSSL mostly finds the nearest DoA in the database,



(a) Correct detection percentage.

(b) Mean absolute error.

Fig. 3: MLSSL simulation results for the left-front microphone at 0 dB SNR.

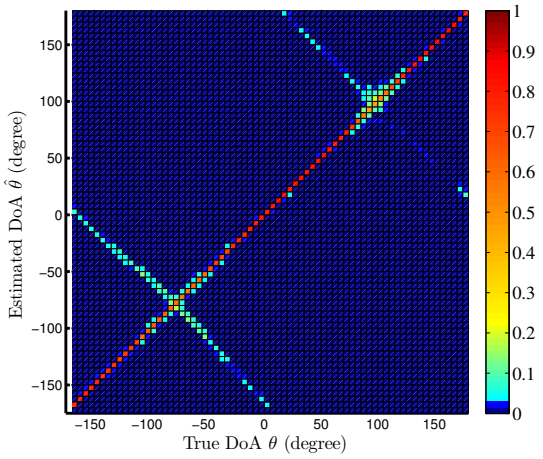


Fig. 4: Confusion matrix of MLSSL for the left-front microphone at 0 dB SNR.

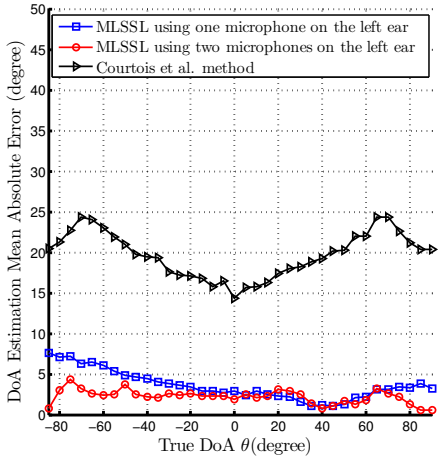


Fig. 5: Performance comparison of MLSSL with the Courtois et al. method for the frontal plane DoAs at the reference SNR of 0 dB.

which means that the resolution of the database is a key factor that influences DoA estimation performance using MLSSL.

5. CONCLUSION AND FUTURE WORK

In this paper, we formulated a target sound DoA estimation problem for a new infrastructure of hearing aid systems, which employs a wireless microphone worn by a sound source of interest. To solve the problem, we considered a maximum likelihood strategy which exploits the noise-free target sound and pre-stored HRTFs. In simulations, MLSSL showed better performance than a recent binaural method proposed by Courtois et al. in [5] even when MLSSL uses only a single microphone. The proposed framework is flexible and easily scalable to any number of microphones. Considering an intelligent search instead of an exhaustive search in the HRTFs database would decrease the computation overhead, and moreover, considering elevation and range in addition to the azimuth will generalize the method. Furthermore, robustness to reverberation is an important issue for SSL. These topics will be investigated in future work.

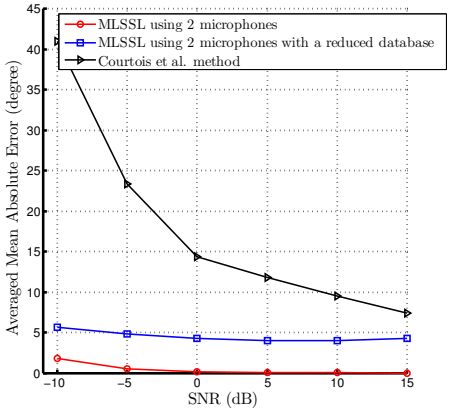


Fig. 6: Mean absolute error averaged over all the DoAs in the front-horizontal plane as a function of SNR.

6. 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 Transactions 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] 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.

[6] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *Proceedings of the 20th European Signal Processing Conference*, 2012, pp. 295–299.

[7] O. Thiergart and E. A. P. Habets, "An informed LCMV filter based on multiple instantaneous direction-of-arrival estimates," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 2013, pp. 659–663.

[8] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, 2001.

[9] D. Hoang, H. F. Silverman, and Y. Ying, "A real-time srp-phat source location implementation using stochastic region contraction(src) on a large-aperture microphone array," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2007, vol. 1, pp. I–121–I–124.

[10] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[11] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space," in *Audio Engineering Society Convention 107*, September 1999.

[12] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, 1997.

[13] Y. Avargel, *Linear System Identification in the Short-Time Fourier Transform Domain*, Ph.D. thesis, Israel Institute of Technology, 2008.

[14] European Hearing Industry Manufactures, "International Speech Test Signal," http://www.ehima.com.

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