ON THE INFLUENCE OF MICROPHONE ARRAY GEOMETRY ON HRTF-BASED SOUND SOURCE LOCALIZATION

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

The direction dependence of Head Related Transfer Functions (HRTFs) forms the basis for HRTF-based Sound Source Localization (SSL) algorithms. In this paper, we show how spectral similarities of the HRTFs of different directions in the horizontal plane influence performance of HRTF-based SSL algorithms; the more similar the HRTFs of different angles to the HRTF of the target angle, the worse the performance. However, we also show how the microphone array geometry can assist in differentiating between the HRTFs of the different angles, thereby improving performance of HRTF-based SSL algorithms. Furthermore, to demonstrate the analysis results, we show the impact of HRTFs similarities and microphone array geometry on an exemplary HRTF-based SSL algorithm, called MLSSL. This algorithm is well-suited for this purpose as it allows to estimate the Direction-of-Arrival (DoA) of the target sound using any number of microphones and any geometries of the microphone array around the head.

Index Terms— Microphone array configuration, Sound source localization, HRTFs, Direction-of-Arrival, Wireless microphone, Hearing Aid Systems.

1. INTRODUCTION

Sound Source Localization (SSL) using a microphone array has been studied in different applications, such as robotics [1, 2, 3], video conferencing [4], and hearing aids [5]. Bio-inspired spatial cues, like Interaural Time Difference (ITD), Interaural Intensity Difference (IID) and the monaural spectral cues in Head Related Transfer Functions (HRTFs) [called Head Related Impulse Responses (HRIRs) in the time domain] are often used for SSL when the microphone array is located next to the ears[1], such as in Hearing Aid Systems (HASs).

Acoustic shadowing effects of the head and torso of a HAS user or a humanoid robot cause the HRTFs to depend on the target sound Direction-of-Arrival (DoA) θ [6]. HRTF-based SSL algorithms use this fact and often exploit a dictionary of HRTFs, labelled by their corresponding θ, to estimate the target sound DoA by finding the best HRTF match in the dictionary [1, 3].

The SSL scenario, which is considered in this paper, is shown in Fig. 1. Because of recent advances in wireless technology for HASs, the depicted scenario is of practical interest. The target signal $s(n)$ is transmitted through the acoustic channel $h_m(n)$ and is "polluted" by environmental noise to generate the noisy signal $r_m(n)$ at microphone m of the HAS. Moreover, we assume that the noise-free target signal $s(n)$ is also available at the HAS via a wireless connection. We aim at estimating θ in this scenario.

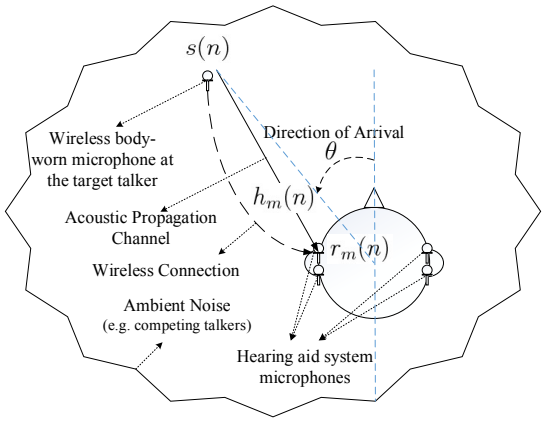


Fig. 1: SSL scenario for a HAS using a wireless microphone: $r_m(n)$, $s(n)$ and $h_m(n)$ are the noisy received sound, the clean target signal and the correspondent HRIR for microphone m, respectively. $s(n)$ is available at the HAS via wireless connection to the wireless microphone. We aim at estimating θ in this scenario.

In general, spectral similarities of the HRTFs and microphone array geometries affect HRTF-based SSL performance. The spectral similarities of the HRTFs in the dictionary may complicate finding the best candidate in the dictionary and reduce the SSL performance. However, the microphone array geometry may assist to improve the SSL performance. Different microphone array geometries impose different amounts of computation and wireless transmission overhead; for example, generally, two different microphone array configurations are conceivable for a HAS: a) a binaural configuration which allows usage of microphones from wirelessly connected hearing aids, but impose wireless transmission overhead, and b) a monaural configuration, which is restricted to use microphones of one hearing aid only, but which does not impose any transmission overhead. The goal of this paper is to compare different microphone array configurations in terms of performance for HRTF-based SSL. Specifically, we wish to study to which extent the need for wireless data transmission in binaural configuration is justified in terms of performance improvements over a monaural configuration.

To demonstrate our investigation results about HRTFs spectral similarities and microphone array geometry, we consider an exemplary SSL algorithm, called Maximum Likelihood Sound Source Localization (MLSSL) [7], that uses the noisy microphone signals, the noise-free target signal and a maximum likelihood (ML) strategy to find the best HRTF match in the dictionary to estimate θ. MLSSL is well-suited for the purpose of this paper since it is scalable to any number of microphones and any array geometry around the head.

[1]While formally, an HRTF is defined to be "a specific individuals left or right ear far-field frequency response, as measured from a specific point in the free field to a specific point in the ear canal" [6], in this paper we use the term HRTF to describe the frequency response from a target source to a microphone of a hearing aid system.

2. SIGNAL MODEL AND MLSSL

In this section, we briefly review the MLSSL algorithm [7]. Regarding Fig. 1, for microphone m of the HAS, we can write:

$$r_m(n) = s(n) * h_m(n) + v_m(n), \qquad m = 1, \cdots, M, \tag{1}$$

where $r_m(n)$, $s(n)$, $h_m(n)$ and $v_m(n)$ are the noisy microphone signal, the noise-free target signal, the HRIR between the target source and microphone m, and the noise signal, respectively. $M \geq 1$ is the number of available HAS microphones, n is the discrete time index, and $*$ represents the convolution operator. It can be shown that Eq. (1) can be approximated in the short-time Fourier-transform (STFT) domain as [7, 8]:

$$R_m(l,k) = S(l,k)H_m(k) + V_m(l,k), \tag{2}$$

where $R_m(l,k)$, $S(l,k)$ and $V_m(l,k)$ are STFT coefficients of the noisy microphone signal, target signal and noise signal for the m^{th} microphone, respectively. $H_m(k)$ is the corresponding HRTF, and l and k are frame and frequency bin indices, respectively.

Collecting expressions for the received microphone signals in a column vector leads to:

$$\boldsymbol{R}(l,k) = S(l,k)\boldsymbol{H}(k) + \boldsymbol{V}(l,k), \tag{3}$$

where

$$\boldsymbol{R}(l,k) = [R_1(l,k), R_2(l,k), \cdots, R_M(l,k)]^{\mathrm{T}}, \tag{4}$$

$$\boldsymbol{H}(k) = [H_1(k), H_2(k), \cdots, H_M(k)]^{\mathrm{T}}, \tag{5}$$

$$\boldsymbol{V}(l,k) = [V_1(l,k), V_2(l,k), \cdots, V_M(l,k)]^{\mathrm{T}}. \tag{6}$$

Assume we possess a dictionary $\mathcal{H} = \{\boldsymbol{H}_1, \boldsymbol{H}_2, \cdots, \boldsymbol{H}_I\}$ of I sets of HRTFs labelled by their corresponding θs, then MLSSL aims at finding the $\boldsymbol{H}_i$ in $\mathcal{H}$ that fits best the observed signals, and in this way estimate the target θ.

Let us assume that $\boldsymbol{V}(l,k)$ in Eq. (3) 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 noise-free $S(l,k)$ is available at the HAS, it is considered as known and deterministic. $\boldsymbol{H}(k)$ is also considered as deterministic but unknown ($\boldsymbol{H} \in \mathcal{H}$). Therefore, $\boldsymbol{R}(l,k)$ in Eq. (3) obeys a Gaussian distribution according to:

$$\boldsymbol{R}(l,k) \sim \mathcal{N}(S(l,k)\boldsymbol{H}(k), \mathbf{C}_V(l,k)). \tag{7}$$

Because $S(l,k)$ is available at the HAS, it is easy to determine the time-frequency regions in the noisy microphones signals where the target speech is essentially absent, and therefore, adaptively estimate $\mathbf{C}_V(l,k)$ over the frames where the noise is dominant. Moreover, for mathematical convenience, the noisy observations are considered to be independent over time and frequencies. Therefore, the likelihood function of $\boldsymbol{H}_i \in \mathcal{H}$ at frame l, regarding the received signals is given by:

$$f_l(\boldsymbol{R}, S; \boldsymbol{H}_i) = \prod_{j=l-\mathrm{D}+1}^{l} \prod_{k=1}^{\mathrm{K}} \frac{1}{\pi^M |\mathbf{C}_V(j,k)|} \mathrm{e}^{\{-\boldsymbol{Z}_i^{\mathrm{H}}(j,k)\mathbf{C}_V^{-1}(j,k)\boldsymbol{Z}_i(j,k)\}}, \tag{8}$$

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 and K are the number of frames and frequency indices, respectively, used for calculating f_l. We assume the target sound source location is fixed during the D frames. The corresponding log-likelihood function is:

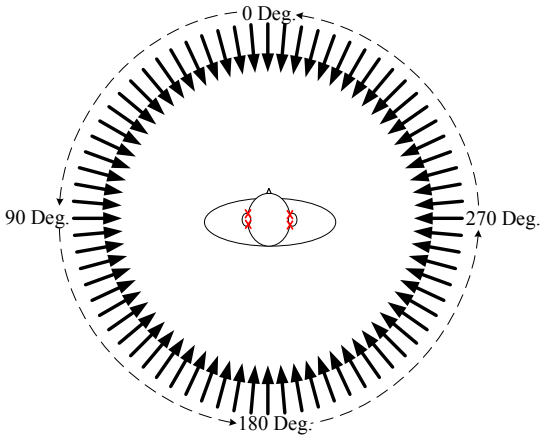


Fig. 2: Acoustic setup. In an anechoic chamber, 72 loudspeakers are placed on a circle with radius of 1.5 m in a horizontal plane centered at the HATS. Possible microphones locations are represented by $\times$ behind the HATS' pinnae.

$$\mathcal{L}_l(\boldsymbol{H}_i) = -M\mathrm{DK}\log\pi - \sum_{j=l-\mathrm{D}+1}^{l}\sum_{k=1}^{\mathrm{K}} \log|\mathbf{C}_V(j,k)| - \sum_{j=l-\mathrm{D}+1}^{l}\sum_{k=1}^{\mathrm{K}} \boldsymbol{Z}_i^{\mathrm{H}}(j,k)\mathbf{C}_V^{-1}(j,k)\boldsymbol{Z}_i(j,k), \tag{9}$$

leading to the maximum likelihood estimation of the HRTF:

$$\boldsymbol{H}_{\mathrm{ML}} = \underset{\boldsymbol{H}_i \in \mathcal{H}}{\arg\max}\ \mathcal{L}_l(\boldsymbol{H}_i) \tag{10}$$

from which the corresponding DoA estimate $\hat{\theta}$ follows. For implementation of Eq. (10), we use an exhaustive search in $\mathcal{H}$.

3. PERFORMANCE ANALYSIS

3.1. Acoustic setup and experiment configurations

For investigating effects of different factors on SSL algorithm performance, an anechoic chamber environment is considered (Fig. 2). The target source can be located at one of 72 uniformly spaced positions, i.e. with 5 degrees resolution, on a horizontal circle with radius 1.5 m centered at a head-and-torso simulator (HATS). Behind-The-Ear (BTE) hearing aids are mounted behind each ear of the HATS. The microphone signals of each hearing aid can be wirelessly exchanged such that a maximum of four microphones can be used to perform SSL. We assume this exchange to be instantaneous and error-free. The distance between front and rear microphones in each hearing aid is 12 mm, and the sampling frequency of the microphone signals is 20 kHz. The STFT uses a frame length of 2048 samples, and a decimation factor of 1024 samples. We use a number of $\mathrm{D} = 2$ frames and the dictionary $\mathcal{H}$ consists of $I = 72$ sets of microphones HRTFs, measured from each loudspeaker to the microphones. The target speech signal is a 10-seconds sample of the ISTS V1.0 [9] which is composed of 21 female voices in 6 different languages.

To generate a realistic and difficult situation, we approximate a cylindrically isotropic large-crowd noise field [10], which is simulated by a number of speech sources that are uniformly spaced on the considered circle. The large-crowd speech signals are from the TSP speech database [11] which consists of different male and female voices. The power of the noise sources is constant for all θs. Therefore, the acoustic shadowing of the HATS causes the effective signal-to-noise-ratios (SNRs) observed at each microphone to be a

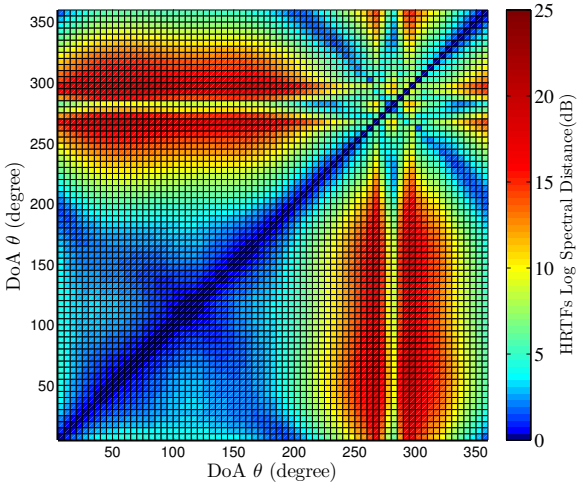


Fig. 3: Log-spectral distances of the HRTFs of θs for the front microphone of the left hearing aid.

function of target direction θ. For this reason, the simulation SNRs are expressed relative to the Left-Front microphone and $\theta = 0^\circ$.

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

$$P_\theta = \frac{Q_\theta}{\mathrm{L}} \times 100, \tag{11}$$

where L is the total frames of the target signal. Furthermore, the mean absolute error (MAE) of the DoA estimation is given by:

$$\sigma_{\hat{\theta}} = \frac{1}{\mathrm{L}} \sum_{j=1}^{\mathrm{L}} |\theta - \hat{\theta}_j|, \tag{12}$$

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

3.2. HRTF similarities

In this section, we study spectral similarities of HRTFs to be able to identify general challenges that any HRTF-based SSL algorithm faces. Intuitively, we expect that HRTF similarities reduce performance of HRTF-based SSL algorithms. To quantify the similarity between two HRTFs H_i and H_j in $\mathcal{H}$, we use the Log-Spectral Distance (LSD) measure [12]:

$$\mathrm{LSD}(H_i, H_j) = \sqrt{\frac{1}{\mathrm{K}} \sum_{k=1}^{\mathrm{K}} \left(20 \log_{10} \frac{|H_i(k)|}{|H_j(k)|} \right)^2}, \tag{13}$$

where $|.|$ denotes the absolute value, and K is the number of frequency bin indices.

Fig. 3 depicts the LSDs of pairs of HRTFs in $\mathcal{H}$ for the front microphone of the left hearing aid. As can be seen, for θs which are on the left side of the head ($\theta \in [0^\circ; 180^\circ]$), i.e. the same side of the hearing aid, their corresponding HRTFs are more similar to each other than to the HRTFs of θs of the other side. This fact helps HRTF-based SSL algorithms to decrease right-left confusions. On the other hand, HRTFs corresponding to angles which are almost symmetric relative to the axis between the left and right ears have similar HRTFs, represented by almost two anti-diagonal blue lines in Fig. 3. These two anti-diagonal blue lines represent the projection of the 3D cone-of-confusion [13] onto the 2D horizontal plane.

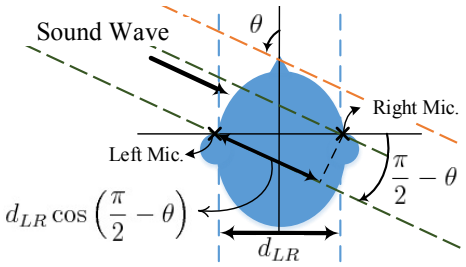


Fig. 4: Left-Right microphone axis.

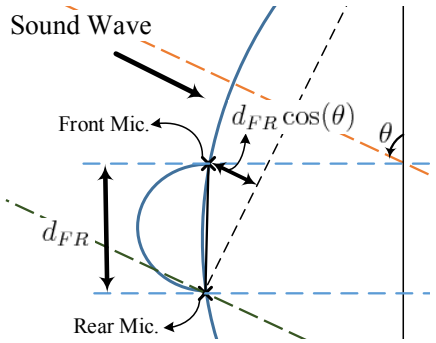


Fig. 5: Front-Rear microphone axis.

These similarities cause front-back confusions, which result in larger estimation errors for the θs in the front or back of the HATS than the left or right sides θs.

3.3. Microphone array configurations

To analyze the impact of microphone array geometry, let us first focus on the two-microphone ($M = 2$) situation. In a HAS context, two configurations are of interest: Left-Right axis (Fig. 4) and Front-Rear axis (Fig. 5). Left-Right axis is a binaural configuration and uses the front microphones of the left and right hearing aids. On the other hand, Front-Rear axis is a monaural configuration and uses the front and rear microphone of a single hearing aid. Without loss of generality, we assume the Front-Rear microphone axis is placed on the left ear. The Left-Right axis needs wireless communication between the hearing aids while the Front-Rear axis does not.

To explain the influence of different configurations on HRTF-based SSL, we analyze the inter-microphone Time Differences of Arrival (TDoA). Since HRTFs can be treated as a minimum phase FIR filter [6], inter-microphone TDoAs are "encoded" in the HRTFs and implicitly affect HRTF-based SSL. To simplify the analysis, we consider a free field and far field situation (ignoring the head and torso filtering effects and assuming a planar wavefront). Let d_{LR} and d_{FR} denote the distance between left and right microphones (Fig. 4), and front and rear microphones (Fig. 5), respectively, and 'c' the sound velocity. The inter-microphone TDoAs for the Left-Right and Front-Rear microphone axes are given by:

$$\tau_{LR} = \frac{d_{LR} \sin\theta}{\mathrm{c}}, \ \tau_{FR} = \frac{d_{FR} \cos\theta}{\mathrm{c}}, \text{respectively.}$$

The different microphone axes provide different sensitivities to changes in θ; the higher the change in TDoA with respect to the change in θ, the better SSL performance. To measure the sensitivity of TDoA to θ changes, the derivatives of τ_{LR} and τ_{FR} with respect to θ are shown in Fig. 6. Clearly, the Front-Rear microphone axis is more sensitive to θ changes when θ is around 90° and 270° while the Left-Right microphone axis is more sensitive to the changes of

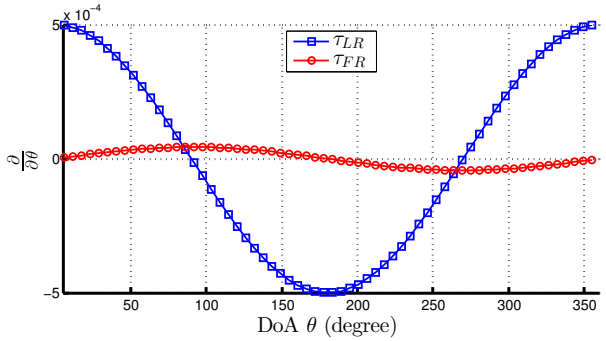
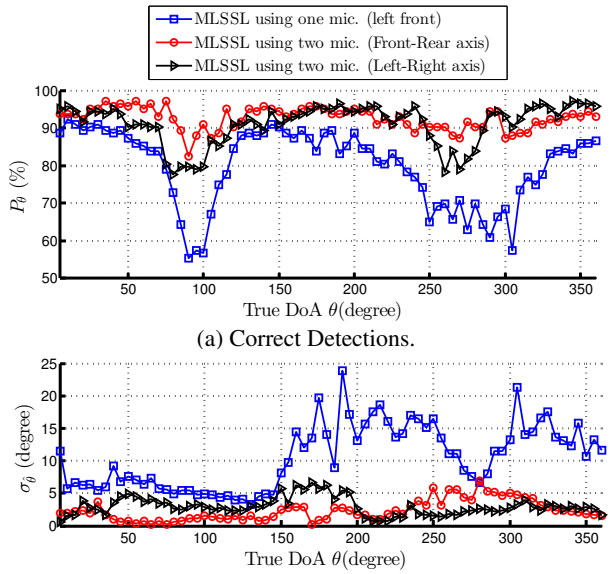


Fig. 6: Inter-Microphone TDoA derivation for different configurations of two microphones.



(a) Correct Detections.

(b) Mean Absolute Error.

Fig. 7: The MLSSL performance using one and two microphones with different axes at 0 dB SNR.

θ when θ is around 0° and 180°. Moreover, the sensitivity to the DoA changes is a function of the microphone distance; the larger the distance, the higher the sensitivity to θ changes.

Regarding Fig. 1, increasing the number of microphones to $M = 3$ enables us to take advantage of both the Left-Right axis and the Front-Rear axis at the cost of computation and wireless transmission overhead. Increasing M further to $M = 4$ will add another Front-Rear axis in the horizontal plane. However, we would not expect this extra Front-Rear axis to provide significant information for SSL in a horizontal plane because the plane is already spanned by the existing microphone axes.

3.4. MLSSL performance

To validate and demonstrate the above analysis, we show the performance of the MLSSL algorithm. Fig. 7 shows P_θ and $\sigma_{\hat{\theta}}$ using one and two microphones signals in different configurations as a function of θ at 0 dB reference SNR. As can be seen in Fig. 7a, P_θ generally falls when the target is located at the sides of the HATS (i.e. $\theta \approx 90^\circ$ and $\theta \approx 270^\circ$), compared with when the target is in the front ($\theta \approx 0$) or behind ($\theta \approx 180^\circ$). This is because the HRTFs around 90° and 270° are locally more similar than the HRTFs around 0° and 180° (Sec. 3.2). Moreover, as can be seen in Fig. 7b, $\sigma_{\hat{\theta}}$ shows different and sometimes opposite behaviour, specifically, for MLSSL using

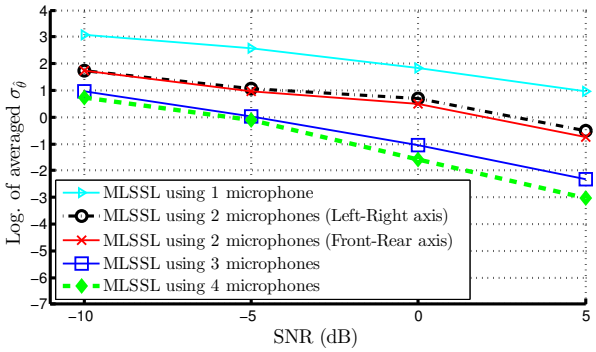


Fig. 8: The MLSSL performance in terms of logarithm of averaged $\sigma_{\hat{\theta}}$ over considered θs for different number of microphones as a function of reference SNR.

one microphone. This behaviour is because of front-back confusions which cause larger estimation errors for the θs in the front or back of the HATS than the left or right sides θs (Sec. 3.2).

Fig. 7 shows that increasing the number of microphones generally improves the performance. However, as expected, the configuration of the microphones also affect MLSSL performance. From Fig. 7a, it is clear that MLSSL ($M = 1$) has lower P_θ for θs around 90° and 270°. For $M = 2$, the Front-Rear configuration is preferred (over the Left-Right) because the Front-Rear axis is more sensitive to changes in θ at these angles (Sec. 3.3, Fig. 6).

Fig. 8 shows the MLSSL performance in terms of the logarithm of the averaged $\sigma_{\hat{\theta}}$ over the 72 θs for different number of microphones as a function of reference SNR. The performance difference between $M = 1$ and $M = 2$ of the MLSSL is significant due to the fact that two microphones can form a new microphone axis in the plane. It is clear that for $M = 2$ the Left-Right axis, which requires wireless communication capabilities, does not offer any advantage over the Front-Rear axis. Increasing the number of microphones to $M = 3$ or $M = 4$, improve the performance of the MLSSL at the cost of higher computation and communication overhead. The performance difference between $M = 2$ and $M = 3$ is also relatively significant, since three microphones configuration allows the MLSSL to make use of both Right-Left and Front-Rear axes via the wireless connection. The performance differences between using $M = 3$ and $M = 4$ are relatively small since the planar dimensions are already spanned when $M = 3$.

4. CONCLUSION

In this paper, we analyzed the performance of HRTF-based SSL algorithms in terms of spectral similarities between HRTFs and microphone array geometry. We showed that due to similarities of different HRTFs, the performance of HRTF-based SSL algorithms depends on the DoA of the target signal. Moreover, we showed that even though increasing the number of microphones in the microphone array improves SSL performance, the geometry of the microphone array plays a key role in improving the performance. For example, a binaural wireless configuration does not necessarily improve the SSL performance compared with a monaural configuration. In this paper, we only considered target locations in the horizontal plane and BTE hearing aids; future research includes considering elevation and range in addition to the azimuth. Furthermore, considering other types of hearing aids than BTE will help complement the investigation.

5. 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] 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*, Sept. 1999.

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

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

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

[10] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, Dec. 2007.

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

[12] A. Gray Jr and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

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