A MACHINE-HEARING SYSTEM EXPLOITING HEAD MOVEMENTS FOR BINAURAL SOUND LOCALISATION IN REVERBERANT CONDITIONS

Ning Ma[⋆]*, Tobias May*[†]*, Hagen Wierstorf*[‡] *and Guy J. Brown*[⋆]

[⋆]Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
[†]Department of Electrical Engineering, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
[‡]AIPA, Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Germany
{n.ma, g.j.brown}@sheffield.ac.uk, tobmay@elektro.dtu.dk, hagen.wierstorf@tu-berlin.de

ABSTRACT

This paper is concerned with machine localisation of multiple active speech sources in reverberant environments using two (binaural) microphones. Such conditions typically present a problem for 'classical' binaural models. Inspired by the human ability to utilise head movements, the current study investigated the influence of different head movement strategies on binaural sound localisation. A machine-hearing system that exploits a multi-step head rotation strategy for sound localisation was found to produce the best performance in simulated reverberant acoustic space. This paper also reports the public release of a free binaural room impulse responses (BRIRs) database that allows the simulation of head rotation used in this study.

Index Terms— head movements, binaural localisation, machine hearing, reverberation, binaural room impulse response

1. INTRODUCTION

The localisation of sound sources in reverberant and noisy environments tends to be challenging for machine systems, but usually presents human listeners with little difficulty. In this paper we investigate whether dynamic features due to head movements can enhance machine localisation. Ultimately, we aim to achieve human-like sound localisation performance in a mobile robot equipped with an anthroporphic dummy head, e.g. the Knowles Electronic Manikin for Acoustic Research (KEMAR). Hence, the current study focuses on auditory-like preprocessing of the acoustic signal, and is constrained to using two (binaural) microphones.

Machine-hearing systems that can reliably estimate the azimuthal position of sound sources are an important building block for a wide range of practical applications, including binaural hearing aids with self-steering beamformers [1], automatic speaker segregation and recognition [2], as well as computational auditory scene analysis (CASA) [3]. By taking an auditory-motivated approach, we anticipate that insights from human hearing will lead to practical advantages in the robustness and power efficiency of the robotic system (e.g., deciding under what conditions the head should be moved, and how far should it be moved).

The cues that listeners use to determine the lateral angle of a sound are interaural time differences (ITDs) and interaural level differences (ILDs). Wallach [4] noted that a particular combination of ITD and ILD cues is not sufficient to define a single location in space. In particular, a given ITD correspond to a number of possible directions on the so-called 'cone of confusion'. In principle, then, listeners might make gross errors when judging location, arising from ambiguity as to whether a sound is located in the front or rear hemifield (i.e., front-back confusions). However, listeners generally hear a sound as originating unambiguously from a single direction. There is ample evidence that listeners use head movements to resolve such ambiguities, by assessing the changes in ITD and ILD as the head is moved [5].

Moreover, listeners may use different types of head movements to resolve ambiguities (rotation in azimuth, tipping, and tilting). Perrett and Noble [6] studied the effect of head movements on sound localisation in three different conditions. Their results showed that for short stimuli of $0.5\,\mathrm{s}$ duration, rotating the head towards the source direction produced significantly fewer front/back errors than free-rotation and when the head was motionless. For long stimuli ($3\,\mathrm{s}$ duration), rotation towards the source direction and free-rotation both produced fewer front/back errors than with short stimuli, but a motionless condition did not. The magnitude of head movements may also vary depending on the task. Wierstorf *et al.* [7] found that listeners tend to make slightly larger head movements when the localization task is most demanding (see also [8]). Similarly, Kim *et al.* [9] found that when listeners were asked to judge sound locations, the maximal rotation azimuth was about 70 degrees, and the average rotation azimuth was about 20 degrees. However, listeners made a wider range of head movements if they were asked to evaluate source width and envelopment.

The current paper makes three main contributions. First, we describe a machine-hearing system that exploits a multi-step head rotation strategy for sound localisation. Our system is evaluated using multiple sound sources in reverberant acoustic spaces, conditions which typically present a problem for 'classical' binaural models. For example, the recent model of Dietz *et al.* [10] was unable to robustly localise multiple sound sources in a small office room ($4\,\mathrm{m}\times\ 5.5\,\mathrm{m}$, $T_{60}\ \sim\ 350\,\mathrm{ms}$). Where head movement has previously been considered in computational systems, the approach has typically been simple (e.g., averaging cross-correlation patterns across different head orientations in order to remove ambiguity), and has assumed anechoic conditions [11]. Secondly, we report the public release of a binaural room impulse response (BRIR) database that allows the simulation of head rotation. Finally, we investigate a number of different head rotation strategies, from which conclusions are drawn about both the underlying principles of head movement in human hearing, and the most practically effective approach for a robotic system.

This work has been supported by the European Union FP7 project TWO!EARS (http://www.twoears.eu) under grant agreement No. 618075.

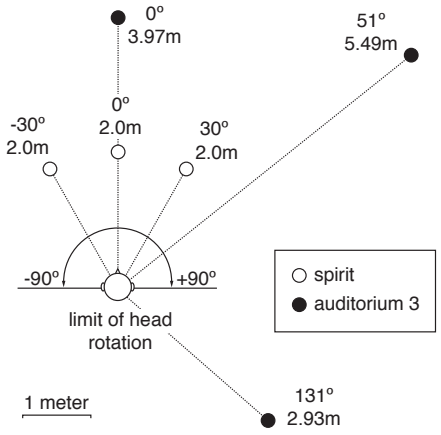

Fig. 1. Position of the sources relative to the head of the listener in the two different rooms.

2. BINAURAL ROOM IMPULSE RESPONSES

The BRIRs were recorded in two different rooms, both of which were located in the TEL building at TU Berlin.[1] The first room was a small office room of size $4.3\,\mathrm{m}$ x $5\,\mathrm{m}$ with a rectangular shape (in the following referred to as room *spirit*). The estimated reverberation time was $T_{60} \sim 0.5\,\mathrm{s}$. The second room was a mid-size lecture room of dimensions $9.3\,\mathrm{m}$ x $9\,\mathrm{m}$ with a trapezium shape and an estimated reverberation time of $T_{60} \sim 0.7\,\mathrm{s}$ (in the following denoted as room *auditorium 3*). The BRIR measurements were conducted for different head orientations ranging from -90 ° to 90 ° with an angular resolution of 1 °. For both rooms, BRIRs for three different source positions were recorded, and their relative positions with respect to the dummy head are illustrated in Fig. 1. The acoustic measurement equipment was exactly the same as described in [12], besides that a Schunk PR-70 Rotary Module was inserted into the KEMAR dummy head in order to perform the head rotations.

3. SYSTEM

3.1. Binaural localisation

Localisation estimates can be made by measuring the interaural time and level differences between the left and the right ear signals, namely ITDs and ILDs. In this study an auditory front-end was employed to analyse ear signals with a bank of 32 Gammatone filters. The centre frequencies of the filters were evenly spaced on the equivalent rectangular bandwidth (ERB) scale between $80\,\mathrm{Hz}$ and $5\,\mathrm{kHz}$. After inner-hair-cell processing, which was approximated by half-wave rectification, the ITDs (based on cross-correlation analysis) and ILDs were estimated for each frequency channel as described in [13] using time frames of $20\,\mathrm{ms}$ duration with $50\,\%$ overlap.

A statistical approach was adopted to learn the mapping between these binaural cues and the corresponding azimuth angle. Specifically, the azimuth- and frequency-dependent distribution of ITDs and ILDs was modelled by a Gaussian mixture model (GMM). To increase the robustness of the system in reverberant multi-source conditions, multi-conditional training (MCT) was applied, where the uncertainties of binaural cues in response to multi-source mixtures are incorporated into the model training [13, 14]. A more detailed description of the binaural localisation system can be found in [15].

[1]The BRIRs are freely available at `http://tinyurl.com/lt76yqs`

3.2. Localisation with head movements

In order to increase the robustness to reverberation, the localisation model is equipped with a hypothesis-driven feedback stage which can trigger a head movement if the source location cannot be unambiguously estimated. The audio inputs are processed on a block-wise basis. The signal of the first block of size T (i.e., frames in the range $t = [1,\ T]$) is used to compute a posterior distribution of the source azimuth using the trained GMMs. In an ideal situation, the local peaks in the posterior distribution correspond to the azimuth of true sources. However, due to the similarities of ITDs and ILDs in the front and in the rear hemifield which can lead to front-back confusions, as well as due to early reflections which can create *phantom sources*, the azimuth posteriors may exhibit more local peaks than the number of actual source positions. Such a hypothesis triggers a head movement in order to solve the localisation confusion. The direction of and the extent of head rotations can be decided by various strategies, which will be discussed in detail in the next section.

Once a head rotation of $\phi\,^\circ$ is completed, a second block of audio will be grabbed and processed in the same way as the first block, but the azimuths will be relative to the new head orientation. Assuming that sources are stationary before and after the head rotation, the two posterior distributions can be used to determine whether any of the local peaks are due to a true source or front-back confusion. If a peak in the first posterior distribution corresponds to a true source position, then it should have moved towards the opposite direction of the head rotation and will appear in the second posterior distribution. On the other hand, if a peak is due to a *phantom source* as a result of front-back confusion, it will not occur at the same position in the second posterior distribution. By exploiting this relationship, potential phantom source peaks are eliminated from both posterior distributions. Finally, the posterior distributions from each block are re-aligned by circular-shifting the azimuth indices by the amount of rotation angles. Since sources are assumed to be stationary, they can be averaged to further emphasise the local peaks corresponding to true sources and cancel out errors. The most prominent peaks in the averaged posterior distribution were assumed to correspond to active source positions. Here the number of active sources was assumed to be known *a prior*. To increase the resolution of the final azimuth estimates, parabolic interpolation was applied to refine the peak positions [16].

3.3. Head rotation strategies

Literature shows that there is a limit to how much listeners move their heads when localising sound sources. For example, in the study by Kim *et al.* [9] the maximal rotation azimuth was found to be around 70 °. Listeners also tend to move the head towards the source position, and there are studies that show head movement towards the sources produces better localisation performance (e.g. [6]). Studying the effect of head rotation strategies in the context of a machine-hearing system is also important for applying such techniques to a robotic platform.

In this paper, the following strategies are evaluated: (1) rotate exactly to the location of the most likely (ML) source; (2) rotate towards the location of the most likely source, but with a fixed rotation angle; (3) random rotation within limits. The most likely source location is decided based on the initial azimuth posterior distribution

before head movement. For the random rotation strategy, a rotation angle in the range $[-90\,^\circ, 90\,^\circ]$ is randomly selected. This range is selected for two reasons: i) the BRIRs used to simulate head rotation were measured with this rotation range; ii) listeners cannot rotate their heads more than $90\,^\circ$.

3.4. Head rotation with multiple steps

Head rotation can either be completed with one step, or with multiple small steps. If a N-step strategy is used, then the signal is divided into $N+1$ blocks in time and the first block is used to choose the overall head rotation angle $\phi\,^\circ$. At each step, a head rotation of a $1/N^{th}$ of $\phi\,^\circ$ is used. This is illustrated in Fig. 2. Such a rotate-stop-listen strategy can be more practical for a robotic platform, as head rotation may produce self-noise which makes the audio collected duration head rotation unusable.

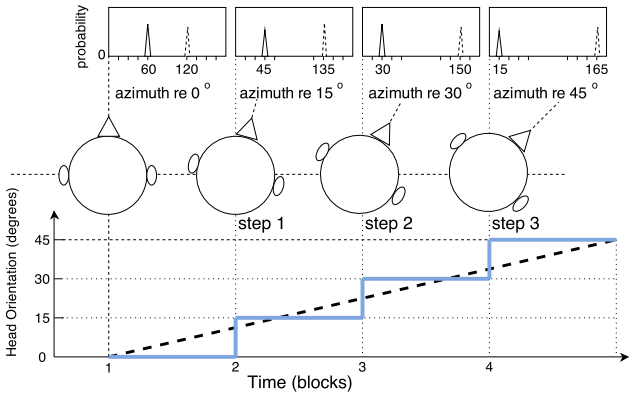

Fig. 2. Illustration of head rotation with small steps. The overall rotation angle is $45\,^\circ$ which is completed with 3 steps as indicated by the solid thick line. The dotted thick line idealises human continues head rotation. Top: azimuth posterior distribution computed for each block re. a different head orientation. The true source azimuth (solid peak) at $60\,^\circ$ moves towards the opposite direction of head rotation while phantom (dotted peak) moves towards the same direction.

4. EXPERIMENTS AND RESULTS

4.1. Experimental setup

In this study sound localisation is not restricted to the frontal plane, and a separate GMM was trained for each of the 72 azimuth angles between -180 ° (left) to 180 ° (right) with an angular resolution of $5\,^\circ$. For each azimuth angle, a 16-component GMM was trained with 20 randomly selected TIMIT sentences [17] that were spatialised using anechoic head related impulse responses (HRIRs) measured with a KEMAR dummy head [18]. To capture the uncertainties due to multiple competing sources and reverberation, the multi-condition training approach in [14] was adopted with diffuse noise (white Gaussian noise) added to each of the training sentences at three different signal-to-noise ratios (SNRs) (20, 10 and 0 dB SNR). For each noisy speech mixture, the binaural features (ITDs and ILDs) used for training included only frames where the *a priori* SNR between the target source and the diffuse background noise exceeded -5 dB.

For evaluation, a set of 100 one-talker, two-talker, and three-talker mixtures was used. Each talker was simulated by randomly selecting a sentence from the TIMIT corpus, excluding the ones used for training. Binaural mixtures of multiple talkers were created by spatialising each talker signal (convolving with the BRIRs described in Section 2) separately before adding them together in each of the two binaural channels. Both rooms contain 3 source positions. For the one-talker and two-talker mixtures, the azimuth directions were randomly selected for a given mixture.

The localisation performance was evaluated by calculating the root mean square error (RMSE) in degrees for each sentence, averaged across all source positions for each room. The MCT-based localisation system described in Sect. 3.1 was selected as a baseline. The proposed localisation system employed the same statistical front-end but adopted various head rotation strategies as described in Section 3.3.

Three experiments were conducted to evaluate the effectiveness of head movements for the proposed computational localisation system. Experiment 1 compared the benefit of different strategies for single head movements. Experiment 2 investigated the influence of the signal duration on localisation performance. Experiment 3 examined the benefit of head rotation with multiple small steps.

4.2. Experiment 1: Effect of different strategies for single movement of the head

For listeners, moving their head to face the source of interest is an optimal strategy, since the minimum audible angle (MAA) is lowest in the front-mid plane [19]. The motivation of this experiment is to identify the optimal strategy for single movement of the head, for the proposed machine-hearing system that employs statistical models of sound localisation.

Short signals of 0.5 s duration were used (TIMIT sentences truncated after 0.5 s onset). The 'No Rotation' baseline integrated source azimuth posterior distributions of each frame (10 ms frame rate) over the entire 0.5-s. For the head rotation systems, the signal was divided into two 0.25-s blocks. The integrated azimuth posterior distribution of the first block was used to decide the head rotation angle.

Table 1 lists the RMSE of the proposed system exploiting various head rotation strategies for localising one, two or three competing talkers in the two rooms. First, all the systems exploiting head rotation improved the localisation accuracy over the 'No Rotation' baseline in both rooms. Second, rotating exactly to the ML source

Table 1. RMSE (in degrees) of systems exploiting a strategy by rotating the head towards the ML source position with a fixed extent for localisation of one, two or three competing talkers in two rooms. 'Random' indicates random head rotation in the range [-90 °, 90 °]. 'ML' indicates rotating exactly to the most likely source direction. The signal duration is 0.5-s.

Rotation strategy	*spirit*			*auditorium 3*		
	Number of competing talkers					
	1	2	3	1	2	3
No Rotation	60	94	91	62	63	60
Random	19	34	53	23	43	57
ML	20	41	63	18	43	53
ML 5 °	24	48	72	15	34	49
ML 10 °	11	41	67	11	35	49
ML 20 °	27	50	61	15	37	52
ML 40 °	10	34	42	11	51	61
ML 60 °	1	23	40	11	40	58
ML 80 °	1	27	40	24	46	56

position did not produce better performance than random rotation. However, the results for the spirit room show that there is benefit to moving the head more ($> 20^\circ$) towards the ML source position.

The best performance was achieved with 60° head movements. The performance is also better than that of random rotation. For the more reverberant *Auditorium 3* room, on the other hand, a greater extent of head rotation did not bring any benefit. This is presumably due to the larger distance of the source positions in the larger auditorium, which effectively reduces the direct-to-reverberant ratio (DRR). As a consequence, the reverberation has a more diffuse character and head movements might not be as beneficial as in the case of the *Spirit* room, in which strong reflections occurred.

4.3. Experiment 2: Effect of signal duration

In this experiment two signal durations were used to measure the effect of signal duration on sound localisation: 0.5-s and 2-s. Three localisation systems were evaluated: the 'No Rotation' baseline, the 'Random Rotation' system and the 'ML-60' system which adopted a strategy of rotating the head towards the ML source position by 60°. Both head rotation systems adopted a single head movement.

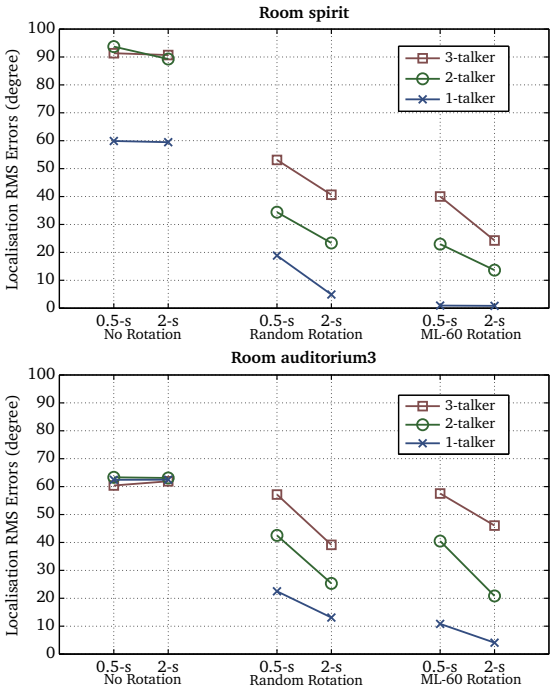

Fig. 3. RMSE in degrees of three localisation systems that exploit either no rotation, a random rotation or a rotation towards the ML source position by 60°. Results are shown for two different signal durations (0.5-s and 2-s).

Figure 3 compares the results produced by the three systems using signals of various durations. Signal duration did not have a strong effect for the 'No Rotation' baseline. If a certain acoustic condition is challenging and produces front-back confusion, it is likely caused by similarities in terms of ITDs and ILDs, which cannot be resolved by integrating the localisation estimates over longer duration. However, both head rotation systems benefitted greatly from having longer signals for localisation. Although integrating localisation estimates over longer duration may not recover the confusion caused by similarities in term of ITDs and ILDs, it could still emphasise the correct source positions and cancel out errors. As a result, the confusions in the more correct azimuth posterior distributions can be better resolved with head rotation. This is also consistent with findings in [6] where longer signal duration has a large benefit on listener's localisation performance in head rotation conditions, but little effect in motionless conditions.

4.4. Experiment 3: Effect of multiple head movements

This experiment evaluated the multi-step head rotation strategy and investigated the trade-off between the number of steps of head rotation that is employed and the time that the system has to integrate binaural cues in between each head movement. The signal length was fixed at 2-s. The best performing ML-60° strategy was used.

Table 2. RMSE in degrees of systems that use a strategy in which the head is rotated towards the ML source position in multiple steps. The overall rotation angle for all conditions was fixed at 60°. The length of all stimuli was 2-s.

Rotation strategy	Angle per step	Block (ms)	*spirit*			*auditorium 3*		
			Number of competing talkers					
			1	2	3	1	2	3
No rotation	0°	2000	59	89	91	62	63	62
1-step	60°	1000	1	14	24	4	21	46
2-step	30°	667	1	25	41	4	34	39
3-step	20°	500	2	21	29	4	20	36
6-step	10°	286	2	17	27	4	21	31
12-step	5°	154	**2**	**6**	**18**	**4**	**13**	**31**

Table 2 lists the localisation RMSE in degrees, as well as rotation angle per step and the time that the system had to integrate binaural cues in between each head movement. In room *spirit*, only the 12-step head rotation system provides a benefit. This could be due to the fact that 'ML-60°' was the best performing strategy for room *spirit* with a single head movement, and the localisation performance is already good. The 12-step head rotation system also produced the best performance in room *auditorium 3*.

5. CONCLUSIONS

In this study, the benefit of different head movement strategies was investigated using a machine-hearing system for binaural sound localisation in reverberant conditions. The performance of localising one to three simultaneous talkers was improved with head rotation over a 'No Rotation' baseline consistently across all the conditions. Using both 0.5-s long and 2-s long signals, the best performing head movement among the tested strategies was to rotate the head towards the most likely source direction. When the duration of signals was increased from 0.5-s to 2-s, the 'No Rotation' baseline produced no improvement while performance of all the head rotation systems was improved. Finally, the system exploiting a multi-step head rotation strategy further improved the localisation performance.

One limitation of the approach is that we only considered lateral movement of the head. An interesting direction for further study is to assess whether changes in the elevation of the head can contribute. The multi-step head rotation can be seen as a first step towards the use of continuous head movements, as exploited by humans. The evidence from these small time segments could be combined using statistical tracking approaches, such as Kalman or particle filters [20].

6. REFERENCES

[1] T. Rohdenburg, S. Goetze, V. Hohmann, K.-D. Kammeyer, and B. Kollmeier, "Objective perceptual quality assessment for self-steering binaural hearing aid microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 2449–2452.

[2] T. May, S. van de Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 2016–2030, 2012.

[3] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Wiley, Hoboken, 2006.

[4] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization," *Journal of Experimental Psychology*, vol. 27, no. 4, pp. 339–368, 1940.

[5] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2841–2853, 1999.

[6] S. Perrett and W. Noble, "The effect of head rotations on vertical plane sound localization," *J. Acoust. Soc. Am.*, vol. 102, no. 4, pp. 2325–2332, 1997.

[7] H. Wierstorf, S. Spors, and A. Raake, "Perception and evaluation of sound fields," in *59th Open Seminar on Acoustics*, 2012.

[8] K. I. McAnally and R. L. Martin, "Sound localization with head movements: Implications for 3D audio displays," *Frontiers in Neuroscience*, vol. 8, no. 210, pp. 1–6, 2014.

[9] C. Kim, R. Mason, and T. Brookes, "Head movements made by listeners in experimental and real-life listening activities," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 425–438, 2013.

[10] M. Dietz, S. D Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Commun.*, vol. 53, no. 5, pp. 592–605, 2011.

[11] J. Braasch, S. Clapp, A. Parks, T. Pastore, and N. Xiang, "A binaural model that analyses acoustic spaces and stereophonic reproduction systems by utilizing head rotations," in *The Technology of Binaural Listening*, J. Blauert, Ed., pp. 201–223. Springer, Berlin, Germany, 2013.

[12] H. Wierstorf, M. Geier, A. Raake, and S. Spors, "A free database of head-related impulse response measurements in the horizontal plane with multiple distances," in *130th AES Conv.*, 2011, p. eBrief 6.

[13] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 1–13, 2011.

[14] J. Woodruff and D. L. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1503–1512, 2012.

[15] T. May, N. Ma, and G. Brown, "Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015.

[16] G. Jacovitti and G. Scarano, "Discrete time techniques for time delay estimation," *IEEE Trans. Signal Process.*, vol. 41, no. 2, pp. 525–533, 1993.

[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic-phonetic continuous speech corpus CD-ROM," *National Inst. Standards and Technol. (NIST)*, 1993.

[18] H. Wierstorf, M. Geier, A. Raake, and S. Spors, "A free database of head-related impulse response measurements in the horizontal plane with multiple distances," in *Proc. 130th Conv. Audio Eng. Soc.*, 2011.

[19] A. W. Mills, "On the minimum audible angle," *J. Acoust. Soc. Amer.*, vol. 30, no. 4, pp. 237–246, 1958.

[20] A. Portello, P. Danès, and S. Argentieri, "Acoustic models and Kalman filtering strategies for active binaural sound localization," in *Proc. Int. Conf. on Intelligent Robots and Systems*, 2011, pp. 137–142.