Exploiting Deep Neural Networks and Head Movements for Robust Binaural Localization of Multiple Sources in Reverberant Environments

Ma, Ning; May, Tobias; Brown, Guy J.

Published in:
IEEE/ACM Transactions on Audio, Speech, and Language Processing

Link to article, DOI:
10.1109/TASLP.2017.2750760

Publication date:
2017

Document Version
Peer reviewed version

Exploiting Deep Neural Networks and Head Movements for Robust Binaural Localization of Multiple Sources in Reverberant Environments

Ning Ma, Tobias May, and Guy J. Brown

***Abstract*—This paper presents a novel machine-hearing system that exploits deep neural networks (DNNs) and head movements for robust binaural localization of multiple sources in reverberant environments. DNNs are used to learn the relationship between the source azimuth and binaural cues, consisting of the complete cross-correlation function (CCF) and interaural level differences (ILDs). In contrast to many previous binaural hearing systems, the proposed approach is not restricted to localization of sound sources in the frontal hemifield. Due to the similarity of binaural cues in the frontal and rear hemifields, front–back confusions often occur. To address this, a head movement strategy is incorporated in the localization model to help reduce the front–back errors. The proposed DNN system is compared to a Gaussian-mixture-model-based system that employs interaural time differences (ITDs) and ILDs as localization features. Our experiments show that the DNN is able to exploit information in the CCF that is not available in the ITD cue, which together with head movements substantially improves localization accuracies under challenging acoustic scenarios, in which multiple talkers and room reverberation are present.**

***Index Terms*—Binaural sound source localisation, deep neural networks, head movements, machine hearing, multi-conditional training, reverberation.**

I. Introduction

THIS paper aims to reduce the gap in performance between human and machine sound localisation, in conditions where multiple sound sources and room reverberation are present. Human listeners have little difficulty in localising sounds under such conditions; they are able to decode the complex acoustic mixture that arrives at each ear with apparent ease [1]. In contrast, sound localisation by machine systems is usually unreliable in the presence of interfering sources and reverberation. This is the case even when an array of multiple microphones is employed [2], as opposed to the two (binaural) sensors available to human listeners.

Manuscript received April 3, 2017; revised July 4, 2017; accepted August 28, 2017. This work was supported by the European Union FP7 project TWO!EARS (http://www.twoears.eu) under Grant 618075. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tuomas Virtanen. *(Corresponding author: Ning Ma.)*

N. Ma and G. J. Brown are with the Department of Computer Science, University of Sheffield, Sheffield S1 4DP, U.K. (e-mail: n.ma@sheffield.ac.uk; g.j.brown@sheffield.ac.uk).

T. May is with the Hearing Systems Group, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark (e-mail: tobmay@elektro.dtu.dk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2017.2750760

The human auditory system determines the azimuth of sounds in the horizontal plane by using two principal cues: interaural time differences (ITDs) and interaural level differences (ILDs). A number of authors have proposed binaural sound localisation systems that use the same approach, by extracting ITDs and ILDs from acoustic recordings made at each ear of an artificial head [3]–[6]. Typically, these systems first use a bank of cochlear filters to split the incoming sound into a number of frequency bands. The ITD and ILD are then estimated in each band, and statistical models such as Gaussian mixture model (GMM) are used to determine the source azimuth from the corresponding binaural cues [6]. Furthermore, the robustness of this approach to varying acoustic conditions can be improved by using multi-conditional training (MCT). This introduces uncertainty into the statistical models of the binaural cues, enabling them to handle the effects of reverberation and interfering sound sources [4]–[7].

In contrast to many previous machine systems, the approach proposed here is not restricted to sound localisation in the frontal hemifield; we consider source positions in the 360° azimuth range around the head. In this unconstrained case, the location of a sound cannot be uniquely determined by ITDs and ILDs; due to the similarity of these cues in the frontal and rear hemifields, front-back confusions occur [8]. Although machine listening studies have noted this as a problem [6], [9], listeners rarely make such confusions because head movements, as well as spectral cues due to the pinnae, play an important role in resolving front-back confusions [8], [10], [11].

Relatively few machine localisation systems have attempted to incorporate head movements. Braasch *et al.* [12] averaged cross-correlation patterns across different head orientations in order to resolve front-back confusions in anechoic conditions. More recently, May *et al.* [6] combined head movements and MCT in a system that achieved robust sound localisation performance in reverberant conditions. In their approach, the localisation system included a hypothesis-driven feedback stage which triggered a head movement when the azimuth could not be unambiguously estimated. Subsequently, Ma *et al.* [9] evaluated the effectiveness of different head movement strategies, using a complex acoustic environment that included multiple sources and room reverberation. In agreement with studies on human sound localisation [13], they found that localisation errors were minimised by a strategy that rotated the head towards the target sound source.

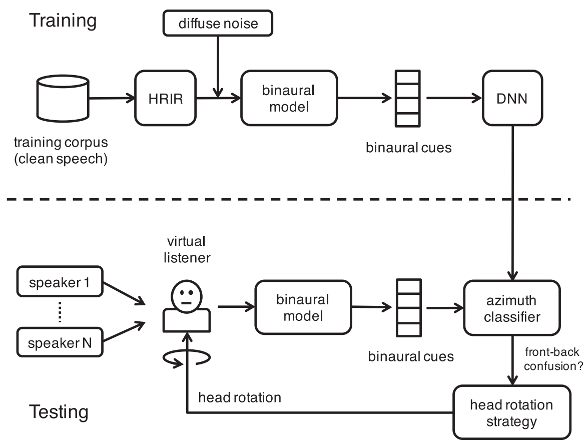

Fig. 1. Schematic diagram of the proposed system, showing steps during training (top) and testing (bottom). During testing, sound mixtures consisting of several talkers are rendered in a virtual acoustic environment, in which a binaural receiver is moved in order to simulate the head rotation of a listener.

This paper describes a novel machine-hearing system that robustly localises multiple talkers in reverberant environments, by combining deep neural network (DNN) classifiers and head movements. Recently, DNNs have been shown to give state-of-the-art performance in a variety of speech recognition and acoustic signal processing tasks [14]. In this study, we use DNNs to map binaural features, obtained from an auditory model, to the corresponding source azimuth. Within each frequency band, a DNN takes as input features the cross-correlation function (CCF) (as opposed to a single estimate of ITD) and the ILD. Using the whole cross-correlation function provides the classifier with rich information for classifying the azimuth of the sound source [15]. A similar approach was used by [16] and [17] in binaural speech segregation systems. However, neither study specifically addressed source localisation because it was assumed that the target source was fixed at zero degrees azimuth.

The proposed binaural sound localisation system is described in detail in Section II. Section III describes the evaluation framework and presents a number of source localisation experiments, in which head movements are simulated by using binaural room impulse responses (BRIRs) to generate direction-dependent binaural sound mixtures. Localisation results are presented in Section IV, which compares our DNN-based approach to a baseline method that uses GMM, and assesses the contribution that various components make to performance. The paper concludes with Section V, which proposes some avenues for future research.

II. System

Figure 1 shows a schematic diagram of the proposed binaural sound localisation system in the full 360° azimuth range. During training, clean speech signals were spatialised using head related impulse responses (HRIRs), and diffuse noise was added before being processed by a binaural model for feature extraction. The noisy binaural features were used to train DNNs to learn the relationship between binaural cues and sound azimuths. During testing, sound mixtures consisting of several talkers are rendered in a virtual acoustic environment, in which a binaural receiver is moved in order to simulate the head rotation of a human listener. The output from the DNN is combined with a head movement strategy to robustly localise multiple talkers in reverberant environments.

A. Binaural Feature Extraction

An auditory front-end was employed to analyse binaural ear signals with a bank of 32 overlapping Gammatone filters, with centre frequencies uniformly spaced on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 8 kHz [18]. Inner-hair-cell processing was approximated by half-wave rectification. No low-pass filtering was employed to simulate the loss of phase-locking at high frequencies as previous studies have shown that in general classifiers are able to exploit the high-frequency structure [4]. Afterwards, the CCF between the right and left ears was computed independently for each frequency band using overlapping frames of 20 ms with a 10 ms shift. The CCF was further normalised by the auto-correlation value at lag zero [4] and evaluated for time lags in the range of $\pm$ 1.1 ms.

Two binaural features, ITDs and ILDs, are typically used in binaural localisation systems [1]. The ITD is estimated as the lag corresponding to the maximum in the CCF. The ILD corresponds to the energy ratio between the left and right ears within the analysis window, expressed in dB. In this study, instead of estimating the ITD the entire CCF was used as localisation features. This approach was motivated by two observations. First, computation of ITDs involves a peak-picking operation which may not be robust in the presence of noise and reverberation. Second, there are systematic changes in the CCF with source azimuth (in particular, changes in the main peak with respect to its side peaks). Even in multi-source scenarios, these can be exploited by a suitable classifier. For signals sampled at 16 kHz, the CCF with a lag range of $\pm$ 1 ms produced a 33-dimensional binaural feature space for each frequency band. This was supplemented by the ILD, forming a final 34-dimensional (34D) feature vector.

B. DNN Localization

DNNs were used to map the 34D binaural feature set to corresponding azimuth angles. A separate DNN was trained for each of the 32 frequency bands. Employing frequency-dependent DNNs was found to be effective for localising simultaneous sound sources. Although simultaneous sources overlap in time, within a local time frame each frequency band is mostly dominated by a single source (Bregman's [19] notion of 'exclusive allocation'). Hence, this allows training using single-source data and removes the need to include multi-source data for training.

The DNN consists of an input layer, two hidden layers, and an output layer. The input layer contained 34 nodes and each node was assumed to be a Gaussian random variable with zero mean and unit variance. The 34D binaural feature inputs for each frequency band were Gaussian normalised, and white Gaussian noise (variance 0.4) was added to avoid overfitting, before being used as input to the DNN. The hidden layers had sigmoid activation functions, and each layer contained 128 hidden nodes. The number of hidden nodes was heuristically selected – more hidden nodes increased the computation time but did not improve localisation accuracy. The output layer contained 72 nodes corresponding to the 72 azimuth angles in the full 360° azimuth range, with a 5° step. A 'softmax' activation function was applied at the output layer. The same DNN architecture was used for all frequency bands and we did not optimise it for individual frequencies.

The neural network was initialised with a single hidden layer, and the number of hidden layers was gradually increased in later training phases. In each training phase, mini-batch gradient descent with a batch size of 128 was used, including a momentum term with the momentum rate set to 0.5. The initial learning rate was set to 1, which gradually decreased to 0.05 after 20 epochs. After the learning rate decreased to 0.05, it was held constant for a further 5 epochs. We also included a validation set and the training procedure was stopped earlier if no new lower error on the validation set could be achieved within the last 5 epochs. At the end of each training phase, an extra hidden layer was added between the last hidden layer and the output layer, and the training phase was repeated until the desired number of hidden layers was reached (two hidden layers in this study).

Given the observed feature set $\boldsymbol{x}_{t,f}$ at time frame t and frequency band f, the 72 'softmax' output values from the DNN for frequency band f were considered as posterior probabilities $\mathcal{P}(k|\boldsymbol{x}_{t,f})$, where k is the azimuth angle and $\sum_k \mathcal{P}(k|\boldsymbol{x}_{t,f}) = 1$. The posteriors were then integrated across frequency to yield the probability of azimuth k, given features of the entire frequency range at time t

$$\mathcal{P}(k|\boldsymbol{x}_t) = \frac{P(k)\prod_f \mathcal{P}(k|\boldsymbol{x}_{t,f})}{\sum_k P(k)\prod_f \mathcal{P}(k|\boldsymbol{x}_{t,f})}, \tag{1}$$

where $P(k)$ is the prior probability of each azimuth k. Assuming no prior knowledge of source positions and equal probabilities for all source directions, Eq. (1) becomes

$$\mathcal{P}(k|\boldsymbol{x}_t) = \frac{\prod_f \mathcal{P}(k|\boldsymbol{x}_{t,f})}{\sum_k \prod_f \mathcal{P}(k|\boldsymbol{x}_{t,f})}. \tag{2}$$

Sound localisation was performed for a signal block consisting of T time frames. Therefore the frame posteriors were further averaged across time to produce a posterior distribution $\mathcal{P}(k)$ of sound source activity

$$\mathcal{P}(k) = \frac{1}{T}\sum_{t}^{t+T-1} \mathcal{P}(k|\boldsymbol{x}_t). \tag{3}$$

The target location was given by the azimuth k that maximised $\mathcal{P}(k)$

$$\hat{k} = \operatorname*{argmax}_k \mathcal{P}(k) \tag{4}$$

C. Localisation With Head Movements

In order to reduce the number of front-back confusions, the proposed localisation model employs a hypothesis-driven feedback stage that triggers a head movement if the source location cannot be unambiguously estimated. A signal block is used to compute an initial posterior distribution of the source azimuth using the trained DNNs. In an ideal situation, the local peaks in the posterior distribution correspond to the azimuths of true sources. However, due to the similarity of binaural features in the front and rear hemifields, *phantom sources* may also become apparent as peaks in the azimuth posterior distribution. Such an ambiguous posterior distribution is shown in the top panel of Fig. 2. In this case, a random head movement within the range of $[-30°, 30°]$ is triggered to solve the localisation confusion. Other possible strategies for head movement are discussed in [9].

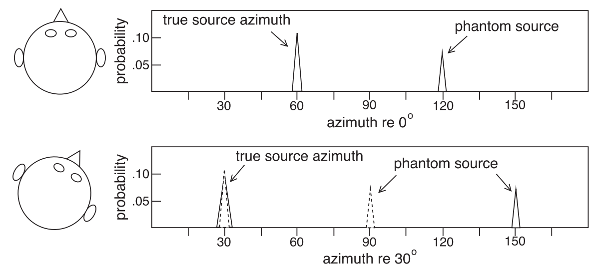

Fig. 2. Illustration of the head movement strategy. Top: posterior probabilities where two candidate azimuths at 60° and 120° are identified. Bottom: after head rotation by 30°, only the azimuth candidate at 30° agrees with the azimuth-shifted candidate from the first signal block (dotted line).

A second posterior distribution is computed for the signal block after the completion of the head movement. If a peak in the first posterior distribution corresponds to a true source position, then it will appear in the second posterior distribution and will be shifted by an amount corresponding to the angle of head rotation (assuming that sources are stationary before and after the head movement). On the other hand, if a peak is due to a phantom source, it will not occur in the second posterior distribution, as shown in the bottom panel of Fig. 2. By exploiting this relationship, potential phantom source peaks are identified and eliminated from both posterior distributions. After the phantom sources have been removed, the two posterior distributions were averaged to further emphasise the local peaks corresponding to true sources. 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 priori*.

The proposed approach to exploiting head movements is based on late information fusion – the information from the model predictions is integrated. This is in contrast to the approach in [12] which adopted early fusion at the feature level by averaging cross-correlation patterns across different head orientations. Late fusion is preferred here for a couple of reasons: i) the use of head rotation is not needed during model training and thus it is more straightforward to generate data for training robust localisation models (DNNs); ii) early feature fusion tends to lose information which can otherwise be exploited by the system. As a result, the proposed system is able to deal with overlapping sound sources in reverberant conditions, while the system reported in [12] was tested in anechoic conditions with a single source.

III. Evaluation

A. Binaural Simulation

Binaural audio signals were created by convolving monaural sounds with HRIRs or BRIRs. For training, an anechoic HRIR

TABLE I
ROOM CHARACTERISTICS OF THE SURREY BRIR DATABASE [21]

	Room A	Room B	Room C	Room D
T_{60} (s)	0.32	0.47	0.68	0.89
DRR (dB)	6.09	5.31	8.82	6.12

catalog based on the Knowles Electronic Manikin for Acoustic Research (KEMAR) head and torso simulator with pinnae [20] was used for simulating the anechoic training signals. The HRIR catalog catalog included impulse responses for the full 360° azimuth range, allowing us to train localisation models for 72 azimuths between 0° and 355° with a 5° step. The models were trained using only the anechoic HRIRs and were not retrained for any room conditions. See Section III-C for more details about training.

For evaluation, the Surrey BRIR database [21] and a BRIR set recorded at TU Berlin [9] were used to reflect different reverberant room conditions. The Surrey database was recorded using a Cortex head and torso simulator (HATS) and includes four room conditions with various amounts of reverberation. The loudspeakers were placed around the HATS on an arc in the median plane, with a 1.5 m radius between ±90° and measured at 5° intervals. Table I lists the reverberation time (T_{60}) and the direct-to-reverberant ratio (DRR) of each room. The anechoic HRIRs used for training were also included to simulate an anechoic condition.

A second set of BRIRs, recorded in the "Auditorium3" room at TU Berlin,[1] was also included particularly for evaluating the benefit of head movements (Section IV-C). The Auditorium3 room is a mid-size lecture room of dimensions 9.3 m × 9 m, with a trapezium shape and an estimated reverberation time T_{60} of 0.7 s. The BRIR measurements were made for different head orientations ranging from -90° to 90° with an angular resolution of 1°. BRIRs for six different source positions, including one in the rear hemifield, were recorded and five of them were selected for this study (two 0° positions are available and the one at 1.5 m away from the head was excluded for simplicity). The five selected source positions with respect to the dummy head are illustrated in Fig. 4.

Note that the anechoic HRIRs used for training and the Surrey BRIRs were recorded using two different dummy heads (KEMAR and Cortex HATS). We use data from two dummy heads because this study is concerned with sound localisation in the 360° azimuth range; the Surrey HATS HRIRs catalog is only available for the frontal azimuth angles and therefore cannot be used to train the full 360° localisation models. However, as the experiment results will show in Section IV, with MCT our proposed systems generalised well despite the HRIR mismatch between training and testing.

Binaural mixtures of multiple competing sources were created by spatialising each source separately at the respective BRIR sampling rate, before adding them together in each of the two binaural channels. In the Auditorium3 BRIRs there is varying distance between the listener position and different source positions. Furthermore there is a difference in impulse response amplitude level even for sources of the equal distance to the listener, likely due to the microphone response difference across recording sessions. To compensate the level difference a scaling factor was computed for each source position by averaging the maximum levels in the impulse responses between left and right ears. The scaling factors were used to adjust the level for each source before spatialisation. As a result the direct sound level of each source when mixed together was approximately the same. For the Surrey BRIR set the level difference did not exist and thus this preprocessing was not applied. The spatialised signals were finally resampled to 16 kHz for training and testing.

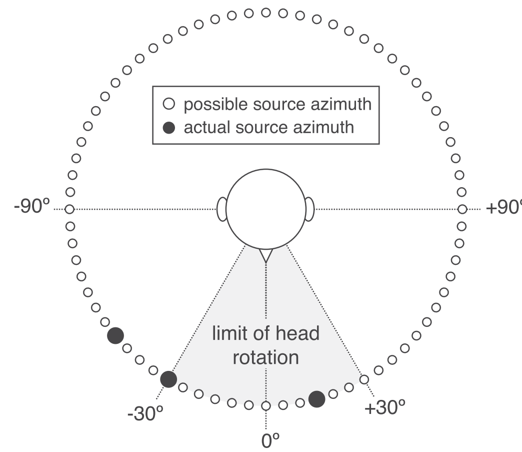

Fig. 3. Schematic diagram of the Surrey BRIR room configuration. Actual source positions were always between ±90°, but the system could report a source azimuth at any of 72 possible azimuths around the head (open circles). Black circles indicate actual source azimuths in a typical three-talker mixture (in this example, at -50°, -30°, and 15°). During testing, head movements were limited to the range $[-30^\circ, 30^\circ]$ as shown by the shaded area.

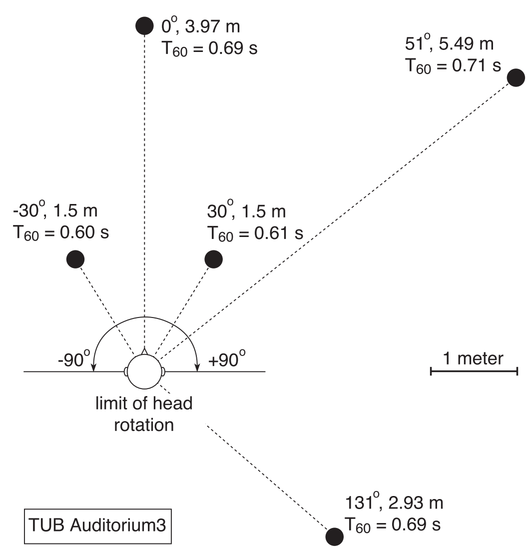

Fig. 4. Schematic diagram of the TUB Auditorium3 configuration. The source distance, azimuth angle and respective T_{60} time are shown for each source.

B. Head Movement Simulation

For the Surrey BRIRs, head movements were simulated by computing source azimuths relative to the head orientation, and

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

loading corresponding BRIRs for the relative source azimuths. Such simulation is only approximate for the reverberant room conditions because the Surrey BRIR database was measured by moving loudspeakers around a fixed dummy head. With the Auditorium3 BRIRs, more realistic head movements were simulated by loading the corresponding BRIR for a desired head orientation. For all experiments, head movements were limited to the range of $\pm 30^{\circ}$.

C. Multi-conditional Training

The proposed systems assumed no prior knowledge of room conditions. The localisation models were trained using only anechoic HRIRs with added diffuse noise, and no reverberant BRIRs were used during training.

Previous studies [4]–[7] have shown that MCT features can increase the robustness of localisation systems in reverberant multi-source conditions. Binaural MCT features were created by mixing a target signal at a specified azimuth with diffuse noise at various signal-to-noise ratios (SNRs). The diffuse noise is the sum of 72 uncorrelated, white Gaussian noise sources, each of which was spatialised across the full 360° azimuth range in steps of 5°. Both the directional target signals and the diffuse noise were created using the same anechoic HRIR recorded using a KEMAR dummy head [20]. This approach was used in preference to adding reverberation during training, since previous studies (e.g., [5]) suggested that it was more likely to generalise well across a wide range of reverberant test conditions.

The training material consisted of speech sentences from the TIMIT database [22]. A set of 30 sentences was randomly selected for each of the 72 azimuth locations. For each spatialised training sentence, the anechoic signal was corrupted with diffuse noise at three SNRs (20, 10 and 0 dB SNR). The corresponding binaural features (ITDs, CCFs, and ILDs) and ILDs) were then extracted. Only those features for which the *a priori* SNR between the target and the diffuse noise exceeded -5 dB were used for training. This negative SNR criterion ensured that the multi-modal clusters in the binaural feature space at higher frequencies, which are caused by periodic ambiguities in the cross-correlation analysis, were properly captured.

D. Experimental Setup

The GRID corpus [23] was used to create three evaluation sets of 50 acoustic mixtures which consisted of one, two or three simultaneous talkers, respectively. Each GRID sentence is approximately 1.5 s long and was spoken by one of 34 native British-English talkers. The sentences were normalised to the same root mean square (RMS) value prior to spatialisation. For the two-talker and three-talker mixtures, the additional azimuth directions were randomly selected from the same azimuth range while ensuring an angular distance of at least 10° between all sources. Each evaluation set included 50 acoustic mixtures which were kept the same for all the evaluated azimuths and room conditions in order to ensure any performance difference was due to test conditions rather than signal variation. Since the duration of each GRID sentence was different, and there was silence of various lengths at the beginning of each sentence, the central 1 s segment of each sentence was selected for evaluation.

Note that although the models were trained and evaluated using speech signals, our systems are not intended to localise only speech sources. Therefore a frequency range from 80 Hz to 8 kHz was selected for the signals sampled at 16 kHz. Our previous studies [6], [15] also show that 32 Gammatone filters (see Section II-A) provide a good tradeoff between frequency resolutions and computational cost. As the evaluation included localisation of up to three overlapping talkers, using too few filters would result in insufficient frequency resolution to reliably localise multiple talkers.

The baseline system was a state-of-the-art localisation system [6] that modelled both ITDs and ILDs features within a GMM framework. As in [6], the GMM modelled the binaural features using 16 Gaussian components and diagonal covariance matrices for each azimuth and each frequency band. The GMM parameters were initialised by 15 iterations of the k-means clustering algorithm and further refined using 5 iterations of the expectation-maximization (EM) algorithm. The second localisation model was the proposed DNN system using the CCF and ILD features. Each DNN employed four layers including two hidden layers each consisting of 128 hidden nodes (see Section II-B).

Both localisation systems were evaluated using different training strategies (clean training and MCT), various localisation feature sets (ITD, ILD and CCF), and with or without head movements. When no head movement was employed, the source azimuths were estimated using the entire 1 s segment from each acoustic mixture. If head movement was used, the 1 s segment was divided into two 0.5 s long blocks and the second block was provided to the system after completion of a head movement. Therefore in both conditions the same signal duration was used for localisation.

The gross accuracy of localisation was measured by comparing true source azimuths with the estimated azimuths. The number of active speech sources N was assumed to be known *a priori* and the N azimuths for which the posterior probabilities were the largest were selected as the estimated azimuths. Localisation of a source was considered accurate if the estimated azimuth was less than or equal to 5° away from the true source azimuth:

$$\mathrm{LocAcc} = \frac{N_{\mathrm{dist}(\phi,\hat{\phi})\leq\theta}}{N} \tag{5}$$

where $\mathrm{dist}(.)$ is the angular distance between two azimuths, ϕ is the true source azimuth, $\hat{\phi}$ is the estimated azimuth, and θ is the threshold in degrees (5° in this study). This metric is preferred to RMS error because our study is concerned with full 360° localisation, and localisation errors in degrees are often large due to front-back confusions.

IV. Results and Discussion

A. Influence of MCT

The first experiment investigated the impact of MCT on the localisation accuracy of the proposed systems. Two scenarios were

TABLE II
GROSS LOCALIZATION ACCURACY IN % FOR VARIOUS SETS OF BRIRS WHEN LOCALIZING ONE, TWO, AND THREE COMPETING TALKERS IN THE FRONTAL HEMIFIELD ONLY AND IN THE FULL 360° RANGE

			Anechoic			Room A			Room B			Room C			Room D			
Hemifiled	Model	MCT	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	Avg.
Frontal	GMM	no	100	99.0	90.5	84.0	63.1	52.8	81.5	59.8	51.8	100	82.5	65.5	88.2	61.2	53.5	75.6
		yes	100	99.9	98.7	99.2	97.1	90.7	100	97.7	91.6	100	99.3	96.5	100	98.4	91.5	97.4
	DNN	no	100	100	99.6	100	99.2	92.2	100	99.0	90.4	100	99.9	96.7	99.9	98.7	91.1	97.8
		yes	100	100	99.7	100	99.5	96.3	100	99.7	96.2	100	99.9	98.2	100	99.6	95.3	99.0
360°	GMM	no	100	97.1	82.6	82.6	48.9	30.7	65.6	38.3	25.3	98.4	70.3	50.2	77.2	46.3	30.0	62.9
		yes	100	100	97.8	99.0	94.2	80.7	97.0	89.0	77.6	100	97.6	88.7	97.3	90.6	79.0	92.6
	DNN	no	100	100	97.4	100	87.0	68.4	94.5	79.0	63.9	97.7	92.5	78.9	94.4	83.4	67.9	87.0
		yes	100	100	98.6	99.7	97.3	87.9	97.2	93.7	86.7	100	97.3	90.2	97.3	94.0	85.0	95.0

The models were trained using either clean training or the MCT method.

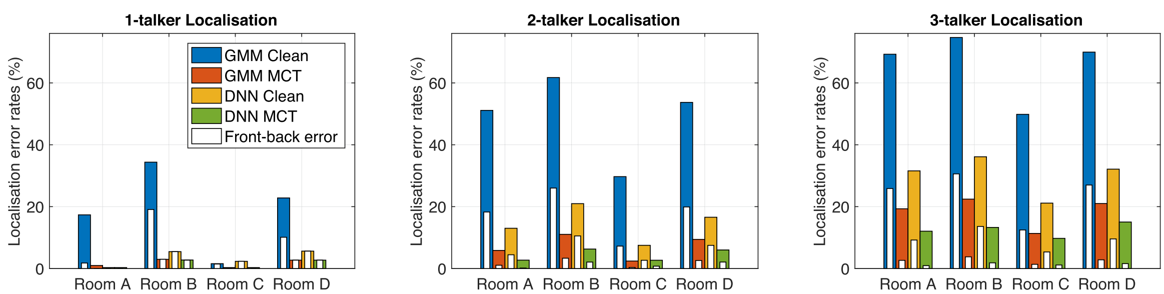

Fig. 5. Localization error rates produced by various systems using either clean training or MCT. Localization was performed in the full 360° range, so that front–back errors could occur, as shown by the white bars for each system. No head movement strategy was employed.

considered: i) sound localisation was restricted to the frontal hemifield so that the systems estimated source azimuths within the range $[-90°, 90°]$; ii) the systems were not informed that the sources lay only in the frontal hemifield and were free to report the azimuth in the full 360° azimuth range. In the second scenario front-back confusions could occur.

Table II lists gross localisation accuracies of all the systems evaluated using various BRIR sets from the Surrey database. First consider the scenario of localisation in the frontal hemifield. For the GMM baseline system, the MCT approach substantially improved the robustness across all conditions, with an average localisation accuracy of 97.4% compared to only 75.6% using clean training. The improvement with MCT was particularly large in multi-talker scenarios and in the presence of room reverberation. For the DNN system, the improvement with MCT over clean training was not as large as that for the GMM system and is only observed in the multi-talker scenarios. The limited improvement is partly because with clean training the performance of the DNN system is already very robust in most conditions, with an average accuracy of 97.8%, which is already better than the GMM system with MCT. This suggests that when localisation was restricted to the frontal hemifield, the DNN can effectively extract cues from the clean CCF-ILD features that are robust in the presence of reverberation.

Considering the case of full 360° localisation, the scenario is more challenging and front-back errors could occur. The GMM system with clean training failed to localise the talkers accurately, with error rates greater than 50% when localising multiple simultaneous talkers. The DNN system with clean training was substantially more robust than the GMM system, but the performance also decreased significantly when multiple talkers were present. The benefit of the MCT method became more apparent for both systems in this scenario – the average localisation accuracy was increased from 62.9% to 92.6% for the GMM system and from 87% to 95% for the DNN system. Across all the room conditions the largest benefits were observed in room B where the direct-to-reverberant ratio was the lowest, and in room D where the reverberation time T_{60} was the longest.

Errors made in 360° localisation could be due to front-back confusion as well as interference caused by reverberation and overlapping talkers. Figure 5 shows errors made by both the GMM and the DNN systems using either clean training or MCT in different room conditions. The errors due to front-back confusions were indicated by white bars for each system. Here a localisation error is considered to be a front-back confusion when the estimated azimuth is within ±20 degrees of the azimuth that would produce the same ITDs in the rear hemifield. It is clear that front-back confusions contributed a large portion of localisation errors for both systems, in particular when clean training was used. When the MCT method was used, not only the errors due to interference of reverberation and overlapping talkers (non-white bar portion in Fig. 5) were greatly reduced, but also the systems produced substantially fewer front-back errors (white bars in Fig. 5). As will be discussed in the next section, without head movements the main cues distinguishing between front-back azimuth pairs lie in the combination of interaural level and time differences (or ITD-related features such as the cross-correlation function). MCT provides the training

TABLE III
GROSS LOCALIZATION ACCURACY IN % USING VARIOUS FEATURE SETS FOR LOCALIZING ONE, TWO, AND THREE COMPETING TALKERS IN THE FULL 360° RANGE

		Anechoic			Room A			Room B			Room C			Room D			
Model	Feature	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	Avg.
GMM	ITD	100	99.8	96.2	99.2	81.6	67.7	91.4	76.6	64.9	97.2	89.4	76.6	89.1	76.6	65.8	84.8
	ITD-ILD	100	100	97.8	99.0	94.2	80.7	97.0	89.0	77.6	100	97.6	88.7	97.3	90.6	79.0	**92.6**
	CCF-ILD	100	100	98.4	100	87.2	73.9	92.1	81.7	71.5	99.9	93.8	81.6	92.6	83.2	72.3	88.5
DNN	CCF	100	100	99.0	99.8	95.8	86.7	91.8	89.5	83.7	98.3	95.8	89.0	91.6	87.8	80.8	92.7
	CCF-ILD	100	100	98.6	99.7	97.3	87.9	97.2	93.7	86.7	100	97.3	90.2	97.3	94.0	85.0	**95.0**

The models were trained using the MCT method. The best feature set for each system is marked in bold font.

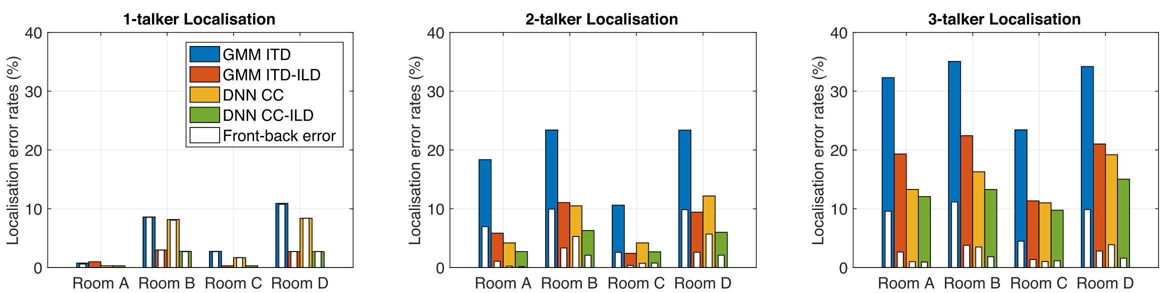

Fig. 6. Comparison of localization error rates produced by various systems using different spatial features. Localization was not restricted in the frontal hemifield so that front–back errors can occur, as indicated by the white bars for each system. No head movement strategy was employed.

stage with better regularisation of the features, which is able to improve the generalisation of the learned models and better discriminate the front-back confusing azimuths.

It is also worth noting that the training and testing stages used HRTFs collected with different dummy heads (the KEMAR was used for training and the HATS was used for testing). However, with MCT the localisation accuracy in the anechoic condition for localising one or two sources was 100%, which suggests that MCT also reduced the sensitivity to mismatches of the receiver.

B. *Contribution of the ILD Cue*

The second experiment investigated the influence of different localisation features, in particular the contribution of the ILD cue. Table III lists the gross localisation accuracies using various feature sets. Here all models were trained using the MCT method and the active head movement strategy was not applied. When ILDs were not used, the GMM performance using just ITDs suffered greatly in reverberant rooms and when localising overlapping talkers; the average localisation accuracy decreased from 92.6% to 84.8%. The performance drop was particularly pronounced in rooms B and D, where the reverberation was strong. For the DNN system, excluding the ILDs also decreased the localisation performance but the performance drop was more moderate, with the average accuracy reduced from 95% to 92.7%. The DNN system using the CCF feature exhibited more robustness in the reverberant multi-talker conditions than the GMM system using the ITD feature. As previously discussed, computation of the ITD involved a peak-picking operation that could be less reliable in challenging conditions, and the systematic changes in the CCF with the source azimuth provided richer information that could be exploited by the DNN.

When ILDs were not used, the localisation errors were largely due to an increased number of front-back errors as suggested by Fig. 6. For single-talker localisation in rooms B and D, without using ILDs almost all the errors made by the systems were front-back errors. When ILDs were used, the number of front-back errors were greatly reduced in all conditions. This suggests that the ILD cue plays a major role in solving the front-back confusions. ITDs or ILDs alone may appear more symmetric between the front and back hemifields, but together with ILDs they create the necessary asymmetries (due to the KEMAR head with pinnae) for the models to learn the differences between front and back azimuths.

Table III also lists localisation results of the GMM system when using the same CCF-ILD feature set as used by the DNN system. The GMM failed to extract the systematic structure in the CCF spanning multiple feature dimensions, most likely due to its inferior ability to model correlated features. The average localisation accuracy is only 88.5% compared to 95% for the DNN system, and again it suffered the most in more reverberant conditions such as rooms B and D.

C. *Benefit of the Head Movement Strategy*

Table IV lists the gross localisation accuracies with or without head movement. All systems were trained using the MCT method and employed the respective best performing features (*GMM ITD-ILD* and *DNN CCF-ILD*).

Both the GMM and DNN systems benefitted from the use of head movements. It is clear from Fig. 7 that the localisation errors were almost entirely due to front-back confusions in one-talker localisation. By exploiting the head movement, the systems managed to reduce most of the front-back errors and achieved near 100% localisation accuracies. In two- or three-talker localisation, the number of front-back errors was also

TABLE IV
GROSS LOCALIZATION ACCURACIES IN % WITH OR WITHOUT THE HEAD MOVEMENT WHEN LOCALIZING ONE, TWO, AND THREE COMPETING TALKERS IN THE FULL 360° AZIMUTH RANGE

Model	Head move	Anechoic			Room A			Room B			Room C			Room D			Avg.
		1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	
GMM	no	100	100	97.8	99.0	94.2	80.7	97.0	89.0	77.6	100	97.6	88.7	97.3	90.6	79.0	92.6
	yes	100	100	97.5	100	97.3	83.4	99.8	93.1	79.9	99.9	99.3	90.8	99.9	93.0	79.5	94.2
DNN	no	100	100	98.6	99.7	97.3	87.9	97.2	93.7	86.7	100	97.3	90.2	97.3	94.0	85.0	95.0
	yes	100	100	98.4	100	99.2	90.0	99.8	96.1	86.9	100	99.0	91.6	99.5	94.7	84.7	96.0

All systems were trained using the MCT method.

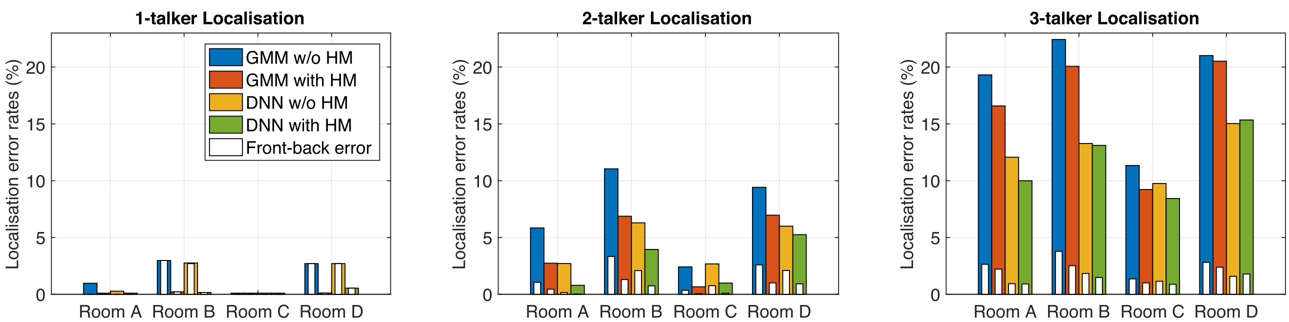

Fig. 7. Localization error rates produced by various systems with or without head movement when localizing one, two, or three overlapping talkers. Localization was performed in the 360° azimuth range so that front–back errors can occur, as indicated by the white bars for each system.

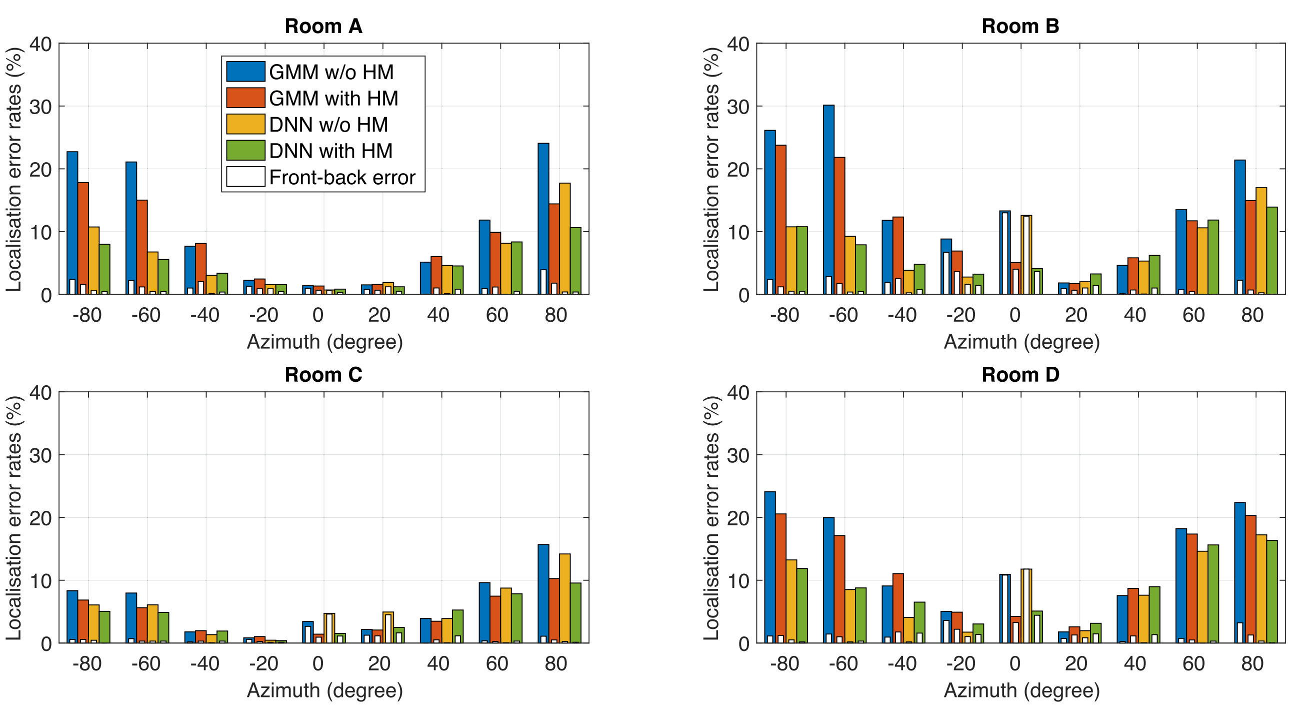

Fig. 8. Localization error rates produced by various systems with or without head movement, as a function of the azimuth. The histogram bin width is 20°. Here the error rates were averaged across the 1-, 2- and 3-talker localization tasks. Localization was performed in the full 360° azimuth range so that front–back errors can occur, as indicated by the white bars for each system.

reduced with the use of head movements. When overlapping talkers were present, the systems produced many localisation errors other than front-back errors, due to the partial evidence available to localise each talker. By removing most front-back errors, the systems were able to further improve the accuracy of localising overlapping sound sources.

Fig. 8 shows the localisation error rates as a function of the azimuth. The error rates here were averaged across the 1-, 2- and 3-talker localisation tasks. Across most room conditions, sound localisation was generally more reliable at more central locations than at lateral source locations. This is particularly the case for the GMM system, as shown in Fig. 8, where the

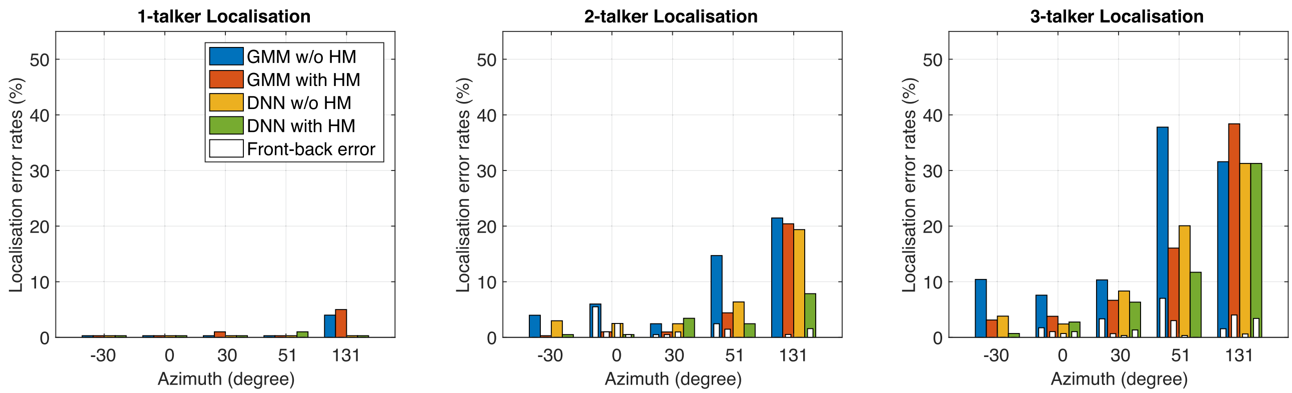

Fig. 9. Localization error rates produced by various systems as a function of the azimuth for the Auditorium3 task. Localization was performed in the full 360° azimuth range so that front–back errors can occur, as indicated by the white bars for each system.

localisation error rates for sources at the sides were above 20% even in the least reverberant Room A. It is also clear from Fig. 8 (white bars) that localisation errors were mostly not due to front-back confusions at lateral azimuths, and in this case the proposed DNN system outperformed the GMM system significantly.

At the central azimuths, on the other hand, almost all the localisation errors were due to front-back confusions. It is noticeable that in more reverberant conditions (such as Rooms B and D), the error rates at the central azimuths $[-10°, 10°]$ were particularly high due to front-back errors for both the GMM and the DNN systems when head movement was not used. The front-back errors were concentrated at central azimuths, probably because binaural features (interaural time and level differences) were less discriminative between 0° and 180° than between the more lateral azimuth pairs.

Finally, Fig. 9 shows the localisation error rates using the Auditorium3 BRIRs in which head movements were more accurately simulated by loading the corresponding BRIR for a given head orientation. Overall the DNN systems significantly outperformed the GMM systems. For single-source localisation the DNN system achieved near 100% localisation accuracy for all source locations including the one at 131° in the rear hemifield. The GMM system produced about 5% error rate for rear source but performed well for the other locations. For two- and three-source localisation, both GMM and DNN systems benefitted from head movements across most azimuth locations. For the GMM system the benefit is particularly pronounced for the source at 51°, with localisation reduced from 14% to 4% in two-source localisation and from 36% to 14% in two-source localisation. The rear source at 131° appeared to be difficult to localise for the GMM system even with head movement, with 20% error rate in two-source localisation. The DNN system with head movements was able to reduce the error rate for the rear source at 131° to 8%.

In general the performance of the models for the 51° and 131° locations is worse than the other source locations when there are multiple sources present at the same time. This is more likely due to the nature of the room acoustics at these locations, e.g., they are further away from the listener and closer to walls. When the sources are overlapping with each other, there are less glimpses left for localisation of each source and with stronger reverberation the sources at 51° and 131° became more difficult to localise.

V. Conclusion

This paper presented a machine-hearing framework that combines DNNs and head movements for robust localisation of multiple sources in reverberant conditions. Since simultaneous talkers were located in a full 360° azimuth range, front-back confusions occurred. Compared to a GMM-based system, the proposed DNN system was able to exploit the rich information provided by the entire CCF, and thus substantially reduced localisation errors. The MCT method was effective in combatting reverberation, and allowed anechoic signals to be used for training a robust localisation model that generalised well to unseen reverberant conditions and to mismatched artificial heads used in training and testing conditions. It was also found that the inclusion of ILDs was necessary for reducing front-back confusions in reverberant rooms. The use of head rotation further increased the robustness of the proposed system, with an average localisation accuracy of 96% under acoustic scenarios where up to three competing talkers and room reverberation were present.

In the current study, the use of DNNs allowed higher-dimensional feature vectors to be exploited for localisation, in comparison with previous studies [4]–[6]. This could be carried further, by exploiting additional context within the DNN either in the time or the frequency dimension. Moreover, it is possible to complement the features used here with other binaural features, e.g., a measure of interaural coherence [24], as well as monaural localisation cues, which are known to be important for judgment of elevation angles [25], [26]. Visual features might also be combined with acoustic features in order to achieve audio-visual source localisation.

The proposed system has been realised in a real world human-robot interaction scenario. The azimuth posterior distributions from the DNN for each processing block were temporally smoothed using a leaky integrator and head rotation was triggered if a front-back confusion was detected in the integrated posterior distribution. Audio signals acquired during head rotation were not processed. Such a scheme can be more practical for a robotic platform as head rotation often produces self-noise which makes the audio unusable.

One limitation of the current systems is that the number of active sources is assumed to be known *a priori*. This can be improved by including a source number estimator that is either learned from the azimuth posterior distribution output by the DNN, or provided directly as an output node in the DNN. The current study only deals with the situation where sound sources are static. Future studies will relax this constraint and address the localisation and tracking of moving sound sources within the DNN framework.

References

[1] J. Blauert, *Spatial Hearing—The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1997.

[2] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1494–1505, Oct. 2014.

[3] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Korner, "A probabilistic model for binaural sound localization," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 5, pp. 982–994, Oct. 2006.

[4] 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, Jan. 2011.

[5] 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, Jul. 2012.

[6] T. May, N. Ma, and G. J. 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, pp. 2679–2683.

[7] T. May, S. van de Par, and A. Kohlrausch, "Binaural localization and detection of speakers in complex acoustic scenes," in *The Technology of Binaural Listening*, J. Blauert, Ed. New York, NY, USA: Springer, 2013, ch. 15, pp. 397–425.

[8] 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.

[9] N. Ma, T. May, H. Wierstorf, and G. J. Brown, "A machine-hearing system exploiting head movements for binaural sound localisation in reverberant conditions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 2699–2703.

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

[11] K. I. McAnally and R. L. Martin, "Sound localization with head movements: Implications for 3D audio displays," *Front. Neurosci.*, vol. 8, pp. 1–6, 2014.

[12] 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. Berlin, Germany: Springer, 2013, pp. 201–223.

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

[14] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[15] N. Ma, G. J. Brown, and T. May, "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," in *Proc. Interspeech*, 2015, pp. 3302–3306.

[16] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2112–2121, Dec. 2014.

[17] Y. Yu, W. Wang, and P. Han, "Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks," *EURASIP J. Audio, Speech, Music Process.*, vol. 2016, no. 1, pp. 1–18, 2016.

[18] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. New York, NY, USA: Wiley/IEEE Press, 2006.

[19] A. Bregman, *Auditory Scene Analysis*. Cambridge, MA, USA: MIT Press, 1990.

[20] 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," *Audio Eng. Soc. Conv. 130*.

[21] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1867–1871, Sep. 2010.

[22] 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," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Internal Rep. 4930, 1993.

[23] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 120, pp. 2421–2424, 2006.

[24] C. Faller and J. Merimaa, "Sound localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, pp. 3075–3089, 2004.

[25] F. Asano, Y. Suzuki, and T. Sone, "Role of spectral cues in median plane localization." *J. Acoust. Soc. Amer.*, vol. 88, no. 1, pp. 159–168, 1990.

[26] P. Zakarauskas and M. S. Cynader, "A computational theory of spectral cue localization," *J. Acoust. Soc. Amer.*, vol. 94, no. 3, pp. 1323–1331, 1993.

Ning Ma obtained the M.Sc. degree with distinction in advanced computer science in 2003 and the Ph.D. degree in hearing-inspired approaches to automatic speech recognition in 2008, both from the University of Sheffield, Sheffield, U.K. He has been a Visiting Research Scientist at the University of Washington, Seattle, WA, USA, and a Research Fellow at the MRC Institute of Hearing Research, Nottingham, U.K., working on auditory scene analysis with cochlear implants. Since 2015, he has been a Research Fellow at the University of Sheffield, working on computational hearing. His research interests include robust automatic speech recognition, computational auditory scene analysis, and hearing impairment. He has authored or coauthored more than 40 papers in these areas.

Tobias May studied hearing technology and audiology and received the M.Sc. degree from the University of Oldenburg, Oldenburg, Germany, in 2007 and the binational Ph.D. degree from the University of Oldenburg in collaboration with the Eindhoven University of Technology, Eindhoven, The Netherlands. Since 2013, he has been with the Department of Electrical Engineering, Technical University of Denmark, first as a Postdoctoral Researcher (2013–2017), and since 2017 as an Assistant Professor. His research interests include computational auditory scene analysis, binaural signal processing, noise-robust speaker identification, and hearing aid processing.

Guy J. Brown received the B.Sc. (Hons.) degree in applied science from Sheffield City Polytechnic, Sheffield, U.K., in 1984, and the Ph.D. degree in computer science from the University of Sheffield, Sheffield, in 1992. He was appointed a Chair of the Department of Computer Science, University of Sheffield, in 2013. He has held visiting appointments at LIMSI-CNRS (France), Ohio State University (USA), Helsinki University of Technology (Finland), and ATR (Japan). He has authored more than 100 papers and is the co-Editor (with Prof. DeLiang Wang) of the IEEE book entitled *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. His research interests include computational auditory scene analysis, speech perception, hearing impairment, and acoustic monitoring for medical applications.