ROBUST LOCALIZATION OF MULTIPLE SPEAKERS EXPLOITING HEAD MOVEMENTS AND MULTI-CONDITIONAL TRAINING OF BINAURAL CUES

Tobias May

Technical University of Denmark
Centre for Applied Hearing Research
DK - 2800 Kgs. Lyngby, Denmark
tobmay@elektro.dtu.dk

Ning Ma, Guy J. Brown

Speech and Hearing Research Group
Department of Computer Science
The University of Sheffield, UK
{n.ma, g.j.brown}@sheffield.ac.uk

ABSTRACT

This paper addresses the problem of localizing multiple competing speakers in the presence of room reverberation, where sound sources can be positioned at any azimuth on the horizontal plane. To reduce the amount of front-back confusions which can occur due to the similarity of interaural time differences (ITDs) and interaural level differences (ILDs) in the front and rear hemifield, a machine hearing system is presented which combines supervised learning of binaural cues using multi-conditional training (MCT) with a head movement strategy. A systematic evaluation showed that this approach substantially reduced the amount of front-back confusions in challenging acoustic scenarios. Moreover, the system was able to generalize to a variety of different acoustic conditions not seen during training.

***Index Terms*—** binaural sound source localization, head movements, multi-conditional training

1. INTRODUCTION

Human sound source localization performance is very robust, even in the presence of multiple competing sounds and room reverberation [1]. The two main cues that are used by the auditory system to determine the azimuth of a sound source are interaural time differences (ITDs) and interaural level differences (ILDs) [2]. However, these binaural cues are not sufficient to uniquely determine the location of a sound [3]. In particular, a given ITD value actually corresponds to a number of possible locations that lie on the so-called *cone of confusion*. Hence, if listeners were only to use these binaural cues, then *front-back confusions* would frequently occur in which a source located in the front hemifield was mistaken for one located in the rear hemifield (or vice versa). In practice, human listeners rarely make front-back confusions because they also use information gleaned from head movements to resolve ambiguities [4, 3, 5].

The long-term aim of the current study is to incorporate human-like binaural sound localisation in a mobile robot with an anthropomorphic dummy head. In a recent paper, we described a software architecture for computational auditory scene analysis (CASA), based on a blackboard system, that incorporates top-down feedback circuits for sensory and motor control [6]. This opens up the possibility of using head movements in a machine hearing system, and the prospect of human-like sound localization performance in challenging acoustic conditions.

Machine hearing systems typically localize sounds by estimating the ITD and ILD in a number of frequency bands, and then mapping these values to an azimuth estimate. Even using static microphones, such approaches can achieve quite promising localization performance. In order to increase the robustness of computational approaches in adverse acoustic conditions, a multi-conditional training (MCT) can be performed, in which the uncertainty of binaural cues in response to multiple sound sources and reverberation is modelled by supervised learning strategies [7, 8, 9]. For example, [9] report gross error rates of less than 5 % for source localization in a variety of reverberant rooms.

Given the good performance of such approaches, the question arises of whether head movements will provide a substantial benefit. However, we note that previous computational approaches have typically been limited to locating sound sources in the frontal hemifield. Hence, although MCT has been shown to provide robust localization performance in the presence of multiple competing sources [7, 8, 9], the learned distribution of binaural cues for sound sources positioned in the front and rear hemifields will be quite similar. It is therefore likely that approaches that use MCT will still suffer from front-back confusions when tested under more demanding conditions. Also, previous work on binaural localization using mobile robots has typically fused information from various positions, but has not used human-like head movements to resolve confusions (e.g., [10]).

The current paper has two aims. First, we describe a machine hearing approach that combines MCT with head movements in order to robustly localize sounds without front-back confusion, while considering the full azimuth range of 360 °. A *virtual listener* is used to verify our approach, in which binaural room impulse responses (BRIRs) are used to spatialise sound sources and simulate head rotation. In our system, a head rotation is requested if the sound source azimuth cannot be unambiguously determined from the estimated ITDs and ILDs. A second aim is to determine whether MCT generalises to different conditions, given that our planned robotic platform may be tested in a variety of acoustic environments and might employ different dummy heads. Specifically, we aim to determine whether a MCT-based sound localisation system can generalise to head related impulse responses (HRIRs) and room acoustics that have not been encountered during training.

2. SYSTEM DESCRIPTION

2.1. Binaural feature extraction

The binaural signals were sampled at a rate of 16 kHz and subsequently analyzed by a bank of 32 Gammatone filters with center

This work was supported by EU FET grant TWO!EARS, ICT-618075.

frequencies equally spaced on the equivalent rectangular bandwidth (ERB) scale between 80 and 5000 Hz [11]. The envelope in each frequency channel was extracted by half-wave rectification. Afterwards, ITDs (based on cross-correlation analysis) and ILDs were estimated according to [7] independently for each frequency channel using overlapping frames of 20 ms duration with a shift of 10 ms. Both binaural cues were combined in a two-dimensional (2D) feature space $\vec{x}_{t,f} = \{\widehat{\mathrm{itd}}_{t,f}, \widehat{\mathrm{ild}}_{t,f}\}$, where t and f denote frame number of frequency channel, respectively.

2.2. GMM-based localization

Sound source localization was performed by a Gaussian mixture model (GMM) classifier that was trained to capture the azimuth- and frequency-dependent distribution of the binaural feature space [7, 8]. Given a set of K sound source directions $\{\varphi_1, \ldots, \varphi_K\}$, that are modeled by frequency-dependent GMMs $\{\lambda_{f,\varphi_1}, \ldots, \lambda_{f,\varphi_K}\}$, a 3D spatial likelihood map can be computed for the kth sound source direction being active at time frame t and frequency channel f

$$\mathcal{L}(t, f, k) = p\left(\vec{x}_{t,f}|\lambda_{f,\varphi_k}\right). \tag{1}$$

The normalized posterior for each frame t was computed by integrating the spatial likelihood map across frequency

$$\mathcal{P}(k|\vec{x}_t) = \frac{P(k)\prod_f \mathcal{L}(t, f, k)}{\sum_{\mathsf{k}} P(\mathsf{k})\prod_f \mathcal{L}(t, f, \mathsf{k})}, \tag{2}$$

where $P(\mathsf{k})$ is the prior probability of each source direction k. Assuming no prior knowledge of source positions and equal probabilities for all source directions, Eq. 2 becomes

$$\mathcal{P}(k|\vec{x}_t) = \frac{\prod_f \mathcal{L}(t, f, k)}{\sum_{\mathsf{k}} \prod_f \mathcal{L}(t, f, \mathsf{k})} \tag{3}$$

To obtain a robust estimation of the sound source azimuth, the frame posteriors were averaged across time for each signal chunk consisting of T time frames to produce a posterior distribution $\mathcal{P}$ of sound source activity

$$\mathcal{P}(k) = \frac{1}{T}\sum_{t}^{t+T-1} \mathcal{P}(k|\vec{x}_t). \tag{4}$$

The most prominent peaks in the posterior distribution $\mathcal{P}$ were assumed to correspond to active source positions. To increase the resolution of the final azimuth estimates, parabolic interpolation was applied to refine the peak positions [12].

2.3. Multi-conditional training

The purpose of MCT is to simulate the uncertainties of binaural cues in response to complex acoustic scenes. This can be achieved by either simulating reverberant BRIRs [7, 8] or by combining HRIRs with diffuse noise [9]. In this study, binaural mixtures were created for the training stage by mixing a target source at a specified azimuth with diffuse noise, which consisted of 72 uncorrelated, white Gaussian noise sources that were placed across the full azimuth range (360 °) in steps of 5 °. The target source was simulated by filtering a randomly selected male or female sentence from the TIMIT database [13] with an anechoic HRIR measured with a Knowles Electronic Manikin for Acoustic Research (KEMAR) dummy head [14]. The same HRIR database was also used for the noise sources.

The localization model was trained with a set of 20 binaural mixtures for each of the 72 azimuth directions. For a given mixture, the target source was corrupted with diffuse noise at three different signal-to-noise ratios (SNRs) (20, 10 and 0 dB SNR), and the corresponding binaural feature space consisting of ITDs and ILDs was extracted. Only those features were used for training, for which the *a priori* SNR between the target and the diffuse noise exceeded -5 dB. 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. In addition, an energy-based voice activity detector (VAD) was used to monitor the activity of the target source. A frame was considered to be silent and excluded from training if the energy level of the target source dropped by more than 40 dB below the global maximum. The resulting binaural feature space was modeled by a GMM classifier with 16 Gaussian components and diagonal covariance matrices for each azimuth and each subband. The corresponding GMM parameters were initialized by 15 iterations of the k-means clustering algorithm and further refined using 5 iterations of the expectation-maximization (EM) algorithm.

In addition to the MCT-based model, a localization model based on *clean* ITDs and ILDs was trained with 20 binaural mixtures which contained the target source only. The feature distribution of the clean binaural feature space was well captured by a GMM with one Gaussian component.

2.4. Head movements

In order to reduce the number of front-back confusions, the localization model is equipped with a hypothesis-driven feedback stage which can trigger a head movement in cases where the azimuth cannot be unambiguously estimated.

Therefore, the first half of the signal chunk (i.e., frames in the range $t = [1,\, T/2]$) is used to derive an initial posterior distribution of the sound source azimuth. If the number of local peaks in the posterior distribution above a pre-defined threshold θ is larger than the number of required source positions, the azimuth information is assumed to be ambiguous, and consequently, a head movement is performed. In this study, we adopted a head movement strategy in which the head is rotated within the range of $[-30\,°, 30\,°]$ in the horizontal plane by a random azimuth degree. If a head movement is triggered, the second half of the signal chunk is re-computed with the new head orientation, and a second posterior distribution is obtained.

Assuming that sources are stationary over the duration of the signal chunk, the initial source azimuth distribution before the head movement can be used to predict the the azimuth distribution after the head movement, given the head rotation azimuth angle. This is done by circular shifting the azimuth indices of the initial azimuth distribution by the amount of the rotation azimuth angle. If a peak in the initial posterior distribution corresponds to a true source posi-

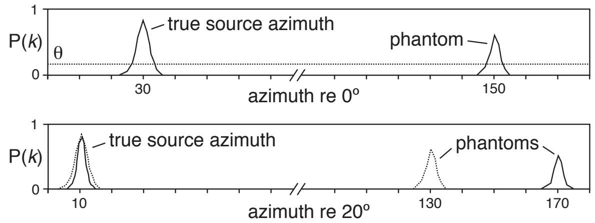

Fig. 1. Head movement strategy. Top: Two candidate azimuths are identified above the threshold θ. Bottom: After head rotation by 20 °, only the azimuth candidate at 10 °agrees with the azimuth-shifted candidate from the first signal block (dotted line).

tion, then it should have moved towards the opposite direction of the head rotation and will appear in the second posterior distribution obtained for the second half of the signal chunk. On the other hand, if a peak is due to 'phantom' sources as a result of front-back confusion, it will not be 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 average of both posterior distributions is taken, giving a final posterior distribution for the signal chunk.

3. EVALUATION

3.1. Binaural listening simulation

In this study, binaural audio signals were created by convolving monaural sounds with HRIRs for anechoic conditions or BRIRs for reverberant conditions. Binaural mixtures of multiple simultaneous sources were created by spatialising each source signal separately before adding them together in each of the two binaural channels.

Two different sets of BRIRs were used to investigate the influence of mismatched binaural recording conditions: i) an anechoic HRIR catalog based on the KEMAR dummy head [14]; ii) the Surrey database [15]. The anechoic KEMAR HRIRs were also used to train the localization models. The Surrey database was captured using a head and torso simulator (HATS) from Cortex Instruments, and includes an anechoic condition as well as four room conditions with various amount of reverberation. The Surrey anechoic condition and the two rooms with the largest T_{60} (room C: $\mathrm{T}_{60} = 0.69\,\mathrm{s}$, $\mathrm{DRR} = 8.82\,\mathrm{dB}$; room D: $\mathrm{T}_{60} = 0.89\,\mathrm{s}$, $\mathrm{DRR} = 6.12\,\mathrm{dB}$) were selected.

Head movements were simulated by computing source azimuths relative to the new head orientation after a head rotation, and loading corresponding HRIRs or BRIRs for the relative source azimuths. This simulation is valid for the two anechoic conditions, in which a head rotation to one direction is equivalent to rotating sources to the opposite direction of the head rotation. The BRIRs of the two room conditions were measured by moving loudspeakers around a fixed dummy head, and thus the simulation is only approximate for the reverberant spaces.

3.2. Experimental setup

The following four localization models were evaluated: (1) a model trained with clean ITDs only, (2) a model trained with clean ITDs and ILDs, (3) a model based on MCT using ITDs and ILDs, and (4) a model based on MCT using ITDs and ILDs, where the binaural feature space consisting of all azimuth angles was normalized to have zero mean and unit variance prior to estimating the GMM parameters. All four localization models were tested with and without the head movement strategy as described in Sect. 2.4. The threshold above which activity in the posterior distribution was considered as source activity was set to $\theta = 0.01$ for all localization models.

All the localization models were tested using a set of 20 one-talker, two-talker, and three-talker acoustic mixtures. During testing, the sound source azimuth was varied in $5\,^\circ$ steps within the range of $[-60\,^\circ, 60\,^\circ]$, as shown in Fig. 2. Source locations were limited to this range of azimuths because the Surrey BRIR database only includes azimuths in the frontal hemifield. However, the system was not provided with information that the azimuth of the source lay within this range, and was free to report the azimuth within the full range of $[-180\,^\circ, 180\,^\circ]$. Hence, front-back confusions could occur if the system incorrectly reported that a source originated from the rear hemifield.

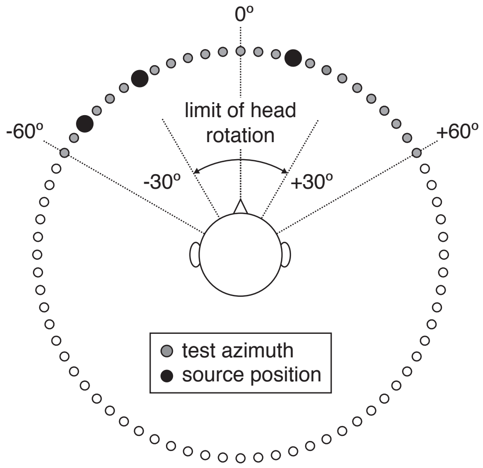

Fig. 2. Schematic diagram of the virtual listener configuration, showing azimuths used for testing (filled circles). Black circles indicate source azimuths in a typical three-talker mixture (in this example, at $-50\,^\circ$, $-30\,^\circ$ and $15\,^\circ$). All azimuths were used for training. During testing, head movements were limited to the range $[-30\,^\circ, 30\,^\circ]$.

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 $10\,^\circ$ between all sources in a mixture. Each talker was simulated by randomly selecting a male or female sentence from the TIMIT corpus, which were different from the ones used for training. The individual sentences were replicated to match the duration of the longest sentence in a given mixture. Each sentence was normalized according to its root mean square (RMS) value prior to spatialization.

The localization performance was evaluated by comparing the true source azimuth with the estimated azimuth obtained from signal chunks of $500\,\mathrm{ms}$ duration. The number of active speech sources was assumed to be known *a priori*. For each binaural mixture, the *gross accuracy* was measured for each signal chunk by counting the number of sources for which the azimuth estimate was within a predefined grace boundary of $\pm 5\,^\circ$. In order to quantify the number of confusions, the quadrant error rate was computed, which was defined as the percentage of azimuth estimates for which the absolute error was greater than $90\,^\circ$.

4. EXPERIMENTAL RESULTS

4.1. Influence of MCT

Localization performance is presented in Tab. 1. When only ITDs were exploited using the clean training data, the azimuth of one speaker was estimated with only $57.7\,\%$ accuracy, which indicates a considerable number of front-back confusions. This confirms that the ITD cue alone is not sufficient to reliably determine the azimuth of a single sound source in anechoic conditions, when considering the full azimuth range of $360\,^\circ$. The joint evaluation of ITDs and ILDs improved performance considerably, which is in line with previous studies [7]. This improvement was particularly noticable for the single-talker mixtures using the anechoic KEMAR recordings.

Table 1. Gross accuracy in % for various sets of BRIRs when localizing one, two and three competing speakers.

Method	Head movement	KEMAR [14]			HATS [15]									Mean
		Anechoic			Anechoic			Room C			Room D			
		1	2	3	1	2	3	1	2	3	1	2	3	
Clean ITD only	No	57.7	22.6	13.5	48.2	22.2	13.6	5.2	3.9	6.2	2.2	1.4	4.9	16.8
	Yes	63.3	25.1	13.9	60.5	25.6	13.6	21.7	6.7	5.8	20.4	5.6	4.6	22.2
Clean	No	91.3	52.2	28.4	65.9	33.9	19.4	26.7	13.0	8.0	13.0	7.6	5.7	30.4
	Yes	99.2	59.3	32.4	69.2	38.8	22.6	64.9	19.7	10.5	64.1	18.9	10.0	42.4
MCT	No	100	88.9	72.4	96.6	81.0	64.1	94.3	62.7	49.1	80.2	48.2	40.7	73.2
	Yes	100	90.0	73.9	97.1	83.0	66.1	99.0	70.7	54.5	95.6	60.5	46.5	78.1
MCT + Norm	No	100	95.5	86.3	100	92.2	82.0	99.7	87.7	76.6	90.6	76.3	68.2	87.9
	Yes	100	96.4	87.7	100	94.8	84.9	99.8	92.5	82.1	97.5	86.3	74.3	91.3

Nevertheless, performance dropped as soon as a different artificial head was used, either in anechoic or reverberant conditions. When using the MCT approach as described in Sect. 2.3, the system was substantially more robust in multi-talker scenarios and in the presence of room reverberation. Also, in contrast to the localization models trained with clean binaural cues, the localization accuracy in anechoic conditions for a single source was 100 % using either the KEMAR or the HATS artificial head, which indicates that the MCT also decreased the sensitivity to mismatches of the receiver. In addition, despite being trained with added white Gaussian noise, the model generalized to recorded BRIRs. This confirms that MCT can account for the distortions of ITDs and ILDs caused by real reverberation. Finally, it can be seen that the feature space normalization provided a large benefit and increased the overall performance by almost 15 %. The normalization stage equalized the range of both ITD and ILD features, which apparently helped to control the weight of the individual GMM components.

4.2. Contribution of head movements

The head movement strategy as described in Sect. 2.4 improved the performance for all localization models. This benefit was particularly pronounced for the single-talker mixtures in the presence of strong reverberation (room C and D), where confusions are likely to occur due to the impact of reflections. Although the model based on clean ITDs and ILDs did not generalize well to the HATS artificial head, the head rotation strategy helped to improve performance in room C and D by more than 40 % for the single-talker scenario. Similarly, head movements were beneficial for the best MCT-based localization model, for which performance increased from 90.6 % to 97.5 % for the most reverberant single-talker scenario.

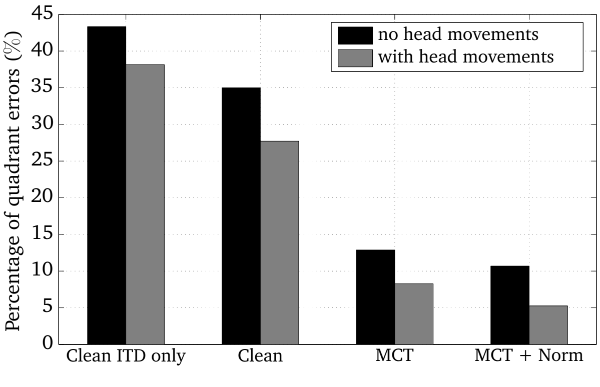

Fig. 3. Percentage of quadrant errors for the four localization models with and without head movements averaged across rooms and the number of speakers.

To quantify the reduction in front-back confusions, the percentage of quadrant errors averaged across all experimental conditions is shown in Fig. 3. It is apparent that the percentage of quadrant errors is systematically reduced, as both ITDs and ILDs are jointly evaluated in combination with the MCT strategy. In particular the MCT strategy substantially reduced the amount of front-back confusions. Nevertheless, there was still a considerable amount of confusion of almost 11 %, which was reduced to 5 % when the MCT-based localization model was combined with the head rotation strategy. This indicates that head rotations provide complementary cues that can be effectively exploited by the localization model to disambiguate sources positioned in the front and in the rear hemifield.

5. DISCUSSION AND CONCLUSION

This paper presented a computational framework that combined the supervised learning of binaural cues with a head rotation strategy, with the aim of robustly estimating the azimuth of multiple speech sources. It was shown that MCT and head movements are complementary, and can be combined to effectively reduce the number of front-back confusions in challenging acoustic scenarios, including multiple competing speakers and reverberation. Furthermore, a systematic evaluation revealed that the system was able to generalize well to unseen acoustic conditions, including a different artificial head that was not used for training.

A simple head movement strategy was considered in the present study, where the rotation angle was randomly chosen and the head orientation was assumed to be stationary across time segments of 250 ms duration. In contrast, humans continuously exploit head movements and also apply different strategies, such as rotating the head towards the source of interest. There is considerable scope for investigating different strategies for head movement in future investigations.

The current approach requires that the number of active speech sources is known. This requirement for *a priori* knowledge could be avoided by blindly estimating the number of active speakers [16]. To enable the localization model to cope with interfering background noise, the framework could also be extended by a source segregation stage, e.g. based on amplitude modulation [17] or pitch [18]. The localization of speakers could subsequently be performed across those segments of contiguous time-frequency units in which speech activity was detected. Finally, the presented localization system should be embedded and tested in a real mobile robot.

6. REFERENCES

[1] M. L. Hawley, R. Y. Litovsky, and H. S. Colburn, "Speech intelligibility and localization in a multi-source environment," *J. Acoust. Soc. Amer.*, vol. 105, no. 6, pp. 3436–3448, 1999.

[2] Jens Blauert, *Spatial hearing - The psychophysics of human sound localization*, The MIT Press, Cambride, MA, USA, 1997.

[3] 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.

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

[6] C. Schymura, N. Ma, T. Walther, G. J. Brown, and D. Kolossa, "Binaural sound source localisation using a bayesian-network-based blackboard system and hypothesis-driven feedback," in *Proc. Forum Acusticum*, Kraków, Poland, 2014.

[7] 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.

[8] 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., chapter 15, pp. 397–425. Springer, Berlin–Heidelberg–New York NY, 2013.

[9] 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.

[10] I. Markovic, A. Portello, P. Danes, I. Petrovic, and S. Argentieri, "Active speaker localization with circular likelihoods and bootstrap filtering," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, Nov 2013, pp. 2914–2920.

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

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

[13] 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.

[14] 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.

[15] 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, 2010.

[16] T. May and S. van de Par, "Blind estimation of the number of speech sources in reverberant multisource scenarios based on binaural signals," in *Proc. IWAENC*, Aachen, Germany, Sep. 2012.

[17] T. May and T. Dau, "Environment-aware ideal binary mask estimation using monaural cues," in *Proc. WASPAA*, 2013, pp. 1–4.

[18] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "A speech fragment approach to localising multiple speakers in reverberant environments," in *Proc. ICASSP*, 2009, pp. 4593–4596.