Proceedings Paper:
Ma, N. orcid.org/0000-0002-4112-3109, Brown, G. orcid.org/0000-0001-8565-5476 and May, T. (Accepted: 2015) Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions. In: Interspeech. INTERSPEECH 2015, 06-10 Sep 2015, Dresden, Germany. International Speech Communication Association , pp. 160-164.

Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions

Ning Ma[1], *Guy J. Brown*[1], *Tobias May*[2]

[1]Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
[2]Hearing Systems Group, Technical University of Denmark, DK - 2800 Kgs. Lyngby, Denmark
{n.ma, g.j.brown}@sheffield.ac.uk, tobmay@elektro.dtu.dk

Abstract

This paper presents a novel machine-hearing system that exploits deep neural networks (DNNs) and head movements for binaural localisation of multiple speakers in reverberant conditions. DNNs are used to map binaural features, consisting of the complete cross-correlation function (CCF) and interaural level differences (ILDs), to the source azimuth. Our approach was evaluated using a localisation task in which sources were located in a full 360-degree azimuth range. As a result, front-back confusions often occurred due to the similarity of binaural features in the front and rear hemifields. To address this, a head movement strategy was incorporated in the DNN-based model to help reduce the front-back errors. Our experiments show that, compared to a system based on a Gaussian mixture model (GMM) classifier, the proposed DNN system substantially reduces localisation errors under challenging acoustic scenarios in which multiple speakers and room reverberation are present.
Index Terms: Binaural source localisation, deep neural networks, head movements, machine hearing, reverberation

1. Introduction

Human listeners usually have little difficulty in localising multiple sound sources in reverberant environments, even though they must decode a complex acoustic mixture arriving at each ear [1]. In contrast, such adverse acoustic environments remain a challenging task for many machine localisation systems, even those employing more than two sensors, such as a microphone array [2].

The auditory system is able to exploit two main cues to determine the azimuth of a sound source in the horizontal plane: interaural time differences (ITDs) and interaural level differences (ILDs). Based on similar principles, binaural sound localisation systems using the ear signals of an artificial head have shown promising localisation performance [3, 4, 5, 6]. Such machine systems typically localise sounds by estimating the ITD and ILD in a number of frequency bands, and employing statistical models such as Gaussian mixture models (GMMs) to map binaural cues to corresponding sound source azimuths. In order to increase the robustness of binaural localisation systems in adverse conditions, multi-conditional training (MCT) can be performed. This introduces uncertainty of binaural cues into the statistical models, allowing them to accommodate to the influence of multiple sound sources and reverberation [4, 5, 6, 7].

Many previous binaural hearing systems have restricted localisation of sound sources to the frontal hemifield. However, if this constraint is relaxed then ITD and ILD cues are often not sufficient to uniquely determine the location of a sound [8]. Due to similarity of these cues in the front and rear hemifields, front-back confusions often occur if sound localisation is performed in the full 360° azimuth range. This problem has been noted in previous machine listening studies, such as [6]. Human listeners, however, rarely make front-back confusions because they also use information gleaned from head movements to resolve ambiguities [9, 8, 10]. This has inspired a few machine localisation systems to incorporate head movement. In [11] cross-correlation patterns were averaged across different head orientations in an attempt to remove front-back ambiguity when localising sounds in anechoic conditions. In [6], head movements were combined with MCT to achieve robust performance of sound localisation in reverberant conditions. The information gleaned from head movements was combined at the statistical model level. In [12], the effectiveness of different head movements was evaluated in a realistic acoustic environment that included multiple speakers and room reverberation. Rotating the head towards the target sound source was found to be the best strategy for minimising localisation errors, an observation that was also found in human sound localisation [13].

This paper presents a novel machine-hearing system that exploits deep neural networks (DNNs) and head movements for robust localisation of multiple speakers in reverberant conditions. DNNs [14] have recently been shown to be very effective classifiers, leading to superior performance in a number of speech recognition and acoustic signal processing tasks. Here, DNNs are used to map binaural features (obtained from a cross-correlogram) to the source azimuth. More specifically, entire cross-correlation functions are used as features (rather than just the time lag of the largest peak) since they provide rich information that can be exploited by the classifier. A similar approach was recently used by [15] for a binaural segregation task. However, their approach assumed that the target source was fixed at zero degrees azimuth, and therefore did not specifically address source localisation.

A binaural sound localisation model that exploits DNNs and head rotations is described in detail in Section 2. Section 3 describes the evaluation framework and presents a number of source localisation experiments. Section 4 presents localisation results and compares our DNN-based approach to a baseline method. Section 5 concludes the paper.

2. System

2.1. 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 [16]. Inner-hair-cell processing was approximated by half-wave rec-

tification. Afterwards, cross-correlation between the right and left ears was computed independently for each frequency channel using overlapping frames of $20\,\mathrm{ms}$ duration with a shift of $10\,\mathrm{ms}$. The cross-correlation function was further normalised by an auto-correlation function at lag zero and evaluated for time lags in the range of $\pm 1.1\,\mathrm{ms}$.

Two features, ITDs and ILDs, are typically used in binaural localisation systems [1]. ITD is estimated as the lag corresponding to the maximum in the cross-correlation function. 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 ITDs, the entire cross-correlation function was used as localisation features. This approach was motivated by two observations. First, computation of the ITD involves a peak-picking operation which may not be robust in the presence of noise. Second, there are systematic changes in the cross-correlation function 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 (see also [17]).

When sampled at 16 kHz, the cross-correlation function with a lag range of ± 1.1 ms produced a 37-dimensional binaural feature space for each frequency channel. This was supplemented by the ILD, forming a final 38-dimensional (38D) feature vector. Similar feature sets were also used in [15] for binaural speech segregation.

2.2. DNN-based localisation

DNNs were used to map the 38D binaural feature set to corresponding azimuth angles. A separate DNN was trained for each frequency channel. The DNN consists of an input layer, 8 hidden layers, and an output layer. The input layer contained 38 nodes and each node was assumed to be a Gaussian random variable with zero mean and unit variance. Therefore the 38D binaural feature input for each frequency channel was first Gaussian normalised, before being fed into the DNN. The hidden layers had sigmoid activation functions, and each layer contained 128 hidden nodes. The number of hidden nodes was heuristically selected as more hidden nodes add more computation and did not improve localisation accuracy in this study. The output layer contained 72 nodes corresponding to the 72 azimuth angles in the full 360° azimuth range (5° steps) considered in this study. The "softmax" activation function was applied at the output layer.

The neural net 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 256 was used, including a momentum term with the momentum rate set to 0.5. The initial learning rate was set to 0.05, which gradually decreased to 0.001 after 10 epochs. After the learning rate decreased to 0.001, it was held constant for a further 5 epochs. At the end of each training phase, an extra hidden layer was added before the output layer, and this training phase was repeated until the desired number of hidden layers was reached (8 hidden layers in this study).

Given the observed feature set $\boldsymbol{x}_{t,f}$ at time frame t and frequency channel f, the 72 "softmax" output values from the DNN for frequency channel 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

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

Sound localisation was performed for a signal chunk 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

$$\mathcal{P}(k) = \frac{1}{T} \sum_{t}^{t+T-1} \mathcal{P}(k|\boldsymbol{x}_t). \quad (2)$$

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

$$\hat{k} = \underset{k}{\operatorname{argmax}}\, \mathcal{P}(k) \quad (3)$$

Previous studies [6, 5, 7] have shown that MCT features can increase the robustness of localisation systems in reverberant multi-source conditions. Here, the DNNs were trained on binaural MCT features created by mixing a target signal at a specified azimuth with diffuse noise at three different signal-to-noise ratios (SNRs) (20 dB, 10 dB and 0 dB). The diffuse noise consisted of 72 uncorrelated, white Gaussian noise sources that were placed across the full azimuth range (360°) in steps of 5°. Both the target signals and the diffuse noise were spatialised by using an anechoic head related impulse response (HRIR) measured with a Knowles Electronic Manikin for Acoustic Research (KEMAR) dummy head [18]. This approach was used in preference to adding reverberation during training, since previous studies (e.g., [5]) suggested that it was likely to give a classifier that performed well across a wide range of reverberant conditions.

2.3. Localisation with head movements

In order to reduce the number of front-back confusions, the DNN localisation model employs a hypothesis-driven feedback stage that triggers a head movement if the source location cannot be unambiguously estimated [12, 6]. A signal chunk 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 azimuth of true sources. However, due to early reflections and the similarity of binaural features in the front and rear hemifields, *phantom sources* may also be apparent as peaks in the azimuth posterior distribution. In this case, a random head movement within the range of $[-30^\circ, 30^\circ]$ is triggered to solve the localisation confusion. Other possible strategies for head movement are discussed in [12].

A second posterior distribution is computed for the signal chunk after the completion of the head movement. Assuming that sources are stationary before and after 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. On the other hand, if a peak is due to a phantom source, it will not occur in the second posterior distribution. 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*.

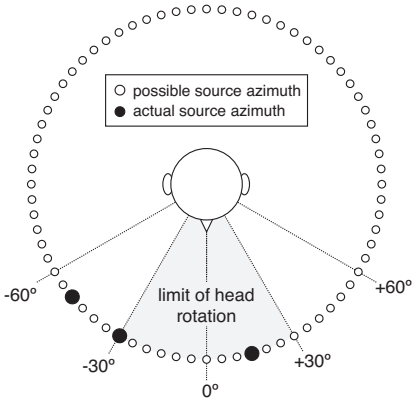

Figure 1: Schematic diagram of the virtual listener configuration. Actual source positions were always between -60° and 60°, but the system could report a source azimuth at any of 72 possible azimuths around the head (open circles). Black circles indicate actual source azimuths in a typical three-talker mixture (in this example, at -50°, -30° and 15°). All azimuths were used for training. During testing, head movements were limited to the range $[-30^\circ, 30^\circ]$ as shown by the shaded area.

3. Evaluation

3.1. Binaural simulation

Binaural audio signals were created by convolving monaural sounds with HRIRs or binaural room impulse responses (BRIRs). An HRIR catalog based on the KEMAR dummy head [18] was used for simulating the anechoic training signals. The evaluation stage used the Surrey BRIR database [19] to simulate reverberant room conditions. The Surrey database was captured using a Cortex head and torso simulator (HATS) and includes four room conditions with various amounts of reverberation. Table 1 lists the reverberation time (T_{60}) and the direct-to-reverberant ratio (DRR) of each room. Binaural mixtures of multiple competing sources were created by spatialising each source signal separately before adding them together in each of the two binaural channels.

Table 1: Details about the room characteristics of the Surrey BRIR database [19] used in this study.

	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

Head movements were simulated by computing source azimuths relative to the head orientation, and 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.

3.2. Experimental setup

During evaluation, the sound source azimuth was varied in 5° steps within the range of $[-60^\circ, 60^\circ]$, as shown in Fig. 1. Source locations were limited to this azimuth range because the Surrey 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.

The GRID corpus [20] was used in this study [1] to form one-talker, two-talker, and three-talker acoustic mixtures. Each GRID sentence is approximately 1.5 s long and of the form "*lay red at G9 now*" 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 in a mixture. Each talker was simulated by randomly selecting sentences from the GRID corpus, which were different from the ones used for training. Each evaluation set included 100 acoustic mixtures.

Three localisation systems were evaluated: i) a baseline system based on GMMs as proposed in [6], which employed both ITDs and ILDs; ii) the proposed DNN system trained without the ILDs (i.e. with only the cross-correlation features); iii) the full DNN system trained with both cross-correlation features and ILDs. The GMM baseline system was trained using the same MCT features. The second system was included in order to determine the role of interaural timing vs. interaural level features in the proposed DNNs.

All three localisation models were tested with and without head movement as described in Section 2.3. When no head movement was used, the source azimuths were estimated from the entire duration of GRID sentences. When head movement was used, a signal chunk of 0.75 s long was taken to compute the first posterior distribution. The rest of the signal from each sentence was taken to compute the second posterior distribution after completion of the head movement.

The localisation performance was evaluated by comparing true source azimuths with the estimated azimuths. The number of active speech sources was assumed to be known *a priori*. For each binaural mixture, the *gross accuracy* was measured for each sentence by counting the number of sources for which the azimuth estimate was within a predefined grace boundary of $\pm 5^\circ$.

4. Results and discussions

Table 2 lists gross localisation accuracy rates of all the systems evaluated for various sets of BRIRs in the Surrey database. When no head movement was exploited, the full DNN system produced substantial improvement over the GMM baseline across all test conditions. The improvement was particularly pronounced in the single-speaker localisation task, with the DNN localisation accuracy approaching 100% in both Room A and Room C. Across all speaker conditions the largest benefits were observed in Room B, where the direct-to-reverberant ratio is the lowest.

When ILDs were not included, however, localisation performance of the DNN system suffered greatly without head movement. The performance drop was particularly pronounced in Room A and B, where even the single-speaker localisation accuracy was below 80% (compared to +90% accuracy for all other systems). Analysis of the types of produced errors sug-

[1] Note our previous studies used the TIMIT corpus [21]. The choice of the corpus, however, did not have much effect on the performance of the binaural localisation systems.

Table 2: Gross accuracy in % for various sets of BRIRs when localising one, two and three competing speakers.

System	Surrey Room A			Surrey Room B			Surrey Room C			Surrey Room D			Mean
	1-spk	2-spk	3-spk	1-spk	2-spk	3-spk	1-spk	2-spk	3-spk	1-spk	2-spk	3-spk	
GMM	92.6	86.3	72.3	87.5	77.6	66.5	92.5	90.5	81.9	92.5	83.4	72.3	83.0
+ Head Movement	99.9	92.1	76.4	99.5	86.4	71.4	99.9	97.8	87.8	99.8	90.0	76.0	89.8
DNN – No ILD	77.0	67.6	63.9	75.2	65.2	62.4	93.8	74.5	69.1	81.6	66.6	61.5	71.5
+ Head Movement	98.6	87.8	73.5	97.7	85.8	71.8	99.8	94.7	81.5	97.5	80.9	67.2	86.4
DNN – Full	99.9	88.7	78.5	94.1	81.5	74.1	100.0	92.2	82.7	97.8	84.9	75.5	87.5
+ Head Movement	99.8	97.1	86.0	99.9	94.9	83.8	100.0	98.4	90.3	99.8	93.7	81.8	93.8

gests that this was largely due to an increased number of front-back errors made by the DNN system when ILDs were not included. Fig. 2 shows the front-back error rates produced by each system with and without head movements in the single-speaker localisation task. It is clear that without head movements, the "DNN – No ILD" system made substantially more front-back errors than the other two systems, especially in Rooms A, B, and D where reverberation was strongest. This suggests the importance of ILDs in resolving front-back confusions in reverberant conditions for the DNN system. Similar observations were also reported for GMM-based localisation systems [4], but the effect was not as detrimental as for the DNN-based system. When the ILDs were included, the front-back errors produced by the DNN system were substantially reduced even without head movements. As Fig. 2 shows, the front-back errors made by the "DNN – Full" system without head movements was close to 0% in all room conditions except Room B.

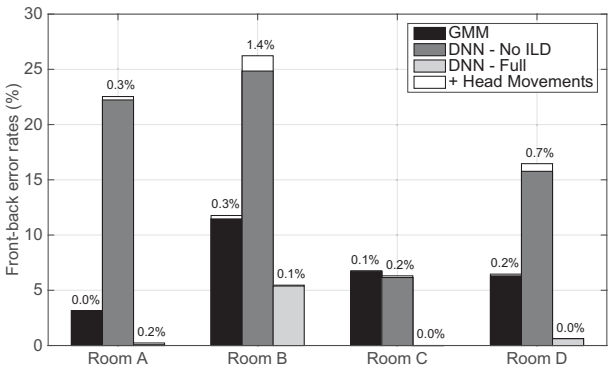

Figure 2: Front-back error rates produced by different systems without head movements (shaded bars) and with head movements (white bars) in different room conditions. The error rate numbers for the white bars (with head movements) are displayed at the top of the bars. The task was single-speaker localisation.

When the head movement strategy was used, the performance of all the systems was considerably improved. As the white bars in Fig. 2 clearly show, the front-back error rates were substantially reduced for all systems. Most systems now made less than 1% front-back errors. The improvement was particularly pronounced for the "DNN – No ILD" system, which reduced the front-back error rates to less than 1.4% in all room reverberations for the single-speaker task. The benefit is clearly translated into overall localisation performance improvement, with the average accuracies for the "DNN – No ILD" system increased from 72% to 86%.

For the "DNN – Full" system the front-back errors were already low for the single-speaker task, and the benefit of head movements was more apparent in the two-speaker and three-speaker localisation conditions. Table 2 shows that in such conditions the improvement due to exploitation of head movements was larger for the DNN-based system than the GMM-based baseline system.

The overall localisation accuracy of the full DNN system is close to 94%, and it consistently outperformed the GMM-based system across all the testing conditions.

5. Conclusions

This paper presented a computational framework that combines deep neural networks and head movements for robust localisation of multiple sources. The DNNs were able to exploit the rich information provided by entire cross-correlation functions. It was also found that including ILDs features produced significantly fewer front-back confusion errors when evaluated in a full 360° azimuth range under challenging acoustic scenarios, in which multiple speakers and room reverberation were present. The use of head rotation further increased the robustness of the proposed DNN-based system, which substantially outperformed a GMM-based baseline system.

In the current study, the use of DNNs allowed higher-dimensional feature vectors to be exploited for localisation, in comparison with previous studies [4, 5, 6]. This could be carried further, by exploiting additional context within the DNN either in the time or frequency dimension. The current study only employed the cross-correlation and ILD features. It is possible to complement the features used here with other binaural features, e.g. a measure of interaural coherence [22], as well as monaural localisation cues, which are known to be important for judgment of elevation angle [23, 24]. Visual features might also be combined with acoustic features in order to achieve audio-visual source localisation.

Finally, a limitation of the current study is that sources were assumed to be static. Future studies will relax this constraint and address the localisation and tracking of moving sound sources within the DNN framework.

6. Acknowledgements

This work was supported by the European Union FP7 project TWO!EARS (http://www.twoears.eu) under grant agreement No. 618075. The authors would like to thank Yulan Liu for helpful discussion on DNN training.

7. References

[1] J. Blauert, *Spatial hearing - The psychophysics of human sound localization*. Cambride, MA, USA: The 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 Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1494–1505, 2014.

[3] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Korner, "A probabilistic model for binaural sound localization," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 36, no. 5, pp. 982–994, 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, 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, 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.

[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. Berlin–Heidelberg–New York NY: 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] 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.

[10] 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.

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

[12] 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. ICASSP*, 2015.

[13] 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.

[14] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

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

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

[17] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, no. 4, pp. 2236–2252, 2003.

[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] 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.

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

[21] 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.

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

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

[24] 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. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/94/3/10.1121/1.408160