forum acusticum 2026
8-12 September · Graz, Austria

COMPARISON OF BINAURAL MODELS FOR THE LOCALIZATION OF VIRTUAL SOUND IMAGES USING A NOVEL TWO-LISTENER VIRTUAL IMAGING SYSTEM

Isaac J. Lambert[1], Vlad S. Paul[1], Philip A. Nelson[1]

[1] *Institute of Sound and Vibration Research, University of Southampton, United Kingdom*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

Virtual sound imaging systems, such as those which work on the principle of crosstalk cancellation (CTC), are often evaluated by means of localization accuracy. For crosstalk cancellation systems, high localization accuracy requires that the actual listener position is the same as the target listener position. In this paper, the inputs to a binaural localization model are varied based on the simulated effect of deviation from target position on a CTC system. The output of the model is compared to localization results from a listening test using a CTC system, revealing that the model is capable of predicting various features of listener responses, including greater variance in prediction for more lateral angles, and asymmetry of variance in left compared to right target azimuths.

Keywords: *crosstalk cancellation, auditory model, localization accuracy*

1. Introduction

Binaural localization modelling is concerned with predicting the perceived direction of sources of sound based on the acoustic pressure at the two ears of a listener. Arguably the first binaural localization model is the Jeffress delay line [1], which extracts interaural time difference (ITD). The human cross-correlator model of Cherry and Sayers [2] introduced temporal integration to the process. A substantial development was the model of Lindemann [3] which introduced contralateral inhibition, and thus sensitivity to interaural level differences (ILD). The model of Gaik [4] introduced further weights to each step of a Lindemann style delay line to further improve sensitivity to ILD. The model of Dietz et al. [5] computes ITDs by means of interaural cross correlation in both the fine structure and envelope of binaural signals, and further utilises the ILD to disambiguate the conversion from interaural phase difference (IPD) to ITD. These ITDs are converted to azimuth estimates by means of a lookup table. The model of May et al. [6] derives localization estimates by learning the feature space relating ITD and ILD for a given source azimuth. Much recent work has applied Bayesian theory to models of sound source localization [7], [8], [9], [10]. The benefits of such an approach include the ability to account for biases in prior assumptions about source position, the ability to model uncertainty in source position, and estimation of source elevation, as well as azimuth. This third ability is a result of incorporating analysis of spectral features of binaural signals, and comparing these to template head-related transfer functions (HRTFs) for different source elevations.

Localization models are particularly useful in the assessment and evaluation of virtual spatial audio algorithms and systems. Due to the considerable time and expense required to conduct listening tests, reliable localization models can provide a more practical method for the evaluation of virtual spatial audio systems. Despite this, little work has attempted to validate the output of localization models against listening test data for virtual audio systems. Wierstorf et al. [11] used an ITD based model to predict localization accuracy for virtual stimuli presented using wave-field synthesis. However, there are substantial differences between wave-field synthesis and CTC. Most notably, CTC systems generally produce a much narrower sweet spot for accurate reproduction, meaning that small changes in listener position can produce large disparities in localization estimates.

This work assesses the ability of existing binaural models of sound localization to predict listener responses in a localization task using a novel two-

***Corresponding author**: ijl1n21@soton.ac.uk.*

listener crosstalk cancellation system. The models exploit ITD and ILD to determine the direction of arrival (DOA) of binaural stimuli. The models were selected for their simplicity and public availability within the Auditory Modeling Toolbox (AMT) [12]. The listening experiment and modelling procedures are outlined in Section 2 (Methods). Section 3 (Results) compares the output of the models to the listening test data. Both models are shown to predict elements of human virtual source localization, with the Dietz model better predicting the standard deviation of listener responses. The May model, in contrast, will be found to better predict the ability of listeners to accurately locate sources from the most lateral target directions ($-90°$ and $90°$).

2. Method

2.1. Listening experiment

Model estimates were compared to listener responses from localization tests conducted by Paul et al. [13] using a novel two-listener virtual imaging system. 25 listeners were presented with samples of pink noise, band-pass filtered between 500 Hz and 10 kHz. The noise was presented from 12 target azimuths, equally spaced between 0° and 330°, with 5 repetitions of each azimuth. Listeners entered the perceived azimuth of the noise sample using a graphical user interface (GUI) in MaxMSP on an iPad. Tests were conducted in the large anechoic chamber at the University of Southampton.

The signal processing employed to render binaural signals using the two-listener virtual imaging system is outlined in Paul et al. [13]. In brief, optimal virtual source strengths $\mathbf{u}_{\text{opt}}$ are calculated as

$$\mathbf{u}_{\text{opt}} = \widetilde{\mathbf{C}}^{\text{H}}(\widetilde{\mathbf{C}}\widetilde{\mathbf{C}}^{\text{H}} + \beta\mathbf{I})^{-1}\mathbf{d}. \quad (1)$$

$\widetilde{\mathbf{C}}$ is the matrix of frequency-domain transfer functions from the virtual sources to the four listener ears. $\mathbf{d}$ is the desired binaural signals, which were computed from the sample of noise ($\mathbf{n}$) using a measured head-related transfer function (HRTF) [14] for the desired source azimuth ($\mathbf{C}_\theta$). The actual source strengths $\mathbf{u}_{\text{opt}}$ were computed from the virtual source strengths as

$$\mathbf{v}_{\text{opt}} = \mathbf{F}\mathbf{u}_{\text{opt}}. \quad (2)$$

$\mathbf{F}$ is a matrix which maps the virtual sources to the actual sources. This includes a crossover network of 4th order Linkwitz-Riley filters which controls the frequency bandwidth of each virtual source.

For a two-listener system, this procedure reproduces desired binaural signal at the ears of two listeners, a left listener and a right listener. In the current work, only the responses of a listener in the left position are considered. Figure 1 shows the array geometry and signal processing employed to calculate the array source inputs.

2.2. Deviation from target listener position

To determine the effect of small listener movements on the binaural inputs, the head-related impulse response for each loudspeaker was measured over a range of dummy head positions. Measured positions consisted of a grid of points, ranging from 10 cm in front to 10 cm behind the target position, and from 20 cm to the left of the target position to 20 cm to the right of the target position. Measurements were taken in increments of 2 cm, meaning that a total of 231 (11×21) positions were measured, with each consisting of 24 (12×2) impulse responses.

2.3. Localization Models

The output of two binaural models were used to predict the localization estimates of listeners. The models were those developed by Dietz et al. [5] and May et al. [6], as implemented in the AMT [12]. Note that (in the case of the May model), certain features of the model may not be implemented exactly as in the original paper. A brief overview of the two models will be provided here, for more details the reader is referred to the original publication of each model.

The peripheral processing in both models is largely similar. Both consist of a gammatone filterbank, followed by halfwave rectification and power law compression. In the case of the Dietz model, the gammatone filterbank is preceeded by bandpass filtering, to model the acoustic influence of the middle ear.

The May model makes use of both the computed ITD and ILD directly for azimuth estimation. For each frequency band and time frame, the azimuth is selected which maximises the log-likelihood of observing the observed binaural feature space (consisting of the ILDs and ITDs across the time frame). The feature spaces for each source azimuth are calculated by means of a Gaussian mixture model. In the AMT, these models are precomputed, and the standard models were used in the current work. In the current work, azimuth predictions were computed for each of 32 frequency channels, spaced linearly on an ERB scale between 80 Hz and 5000 Hz. Azimuth estimates were then pooled across frequency channels by mean averaging. This process was repeated for each 10 ms time frame, returning a total of 56 azimuth estimates per binaural signal.

At the heart of the Dietz model is the separation of fine-structure and envelope cues. Within each frequency band, IPD is calculated for both the fine-structure and envelope of the stimulus. The ILD cue is utilised to disambiguate the conversion from IPD to ITD. This effectively extends the useful upper limit of ITD from 1000 Hz to about 1400 Hz. Within each frequency band, the ITD is converted to an azimuth estimate by means of a lookup table. This table contains (for each frequency band) the ITD

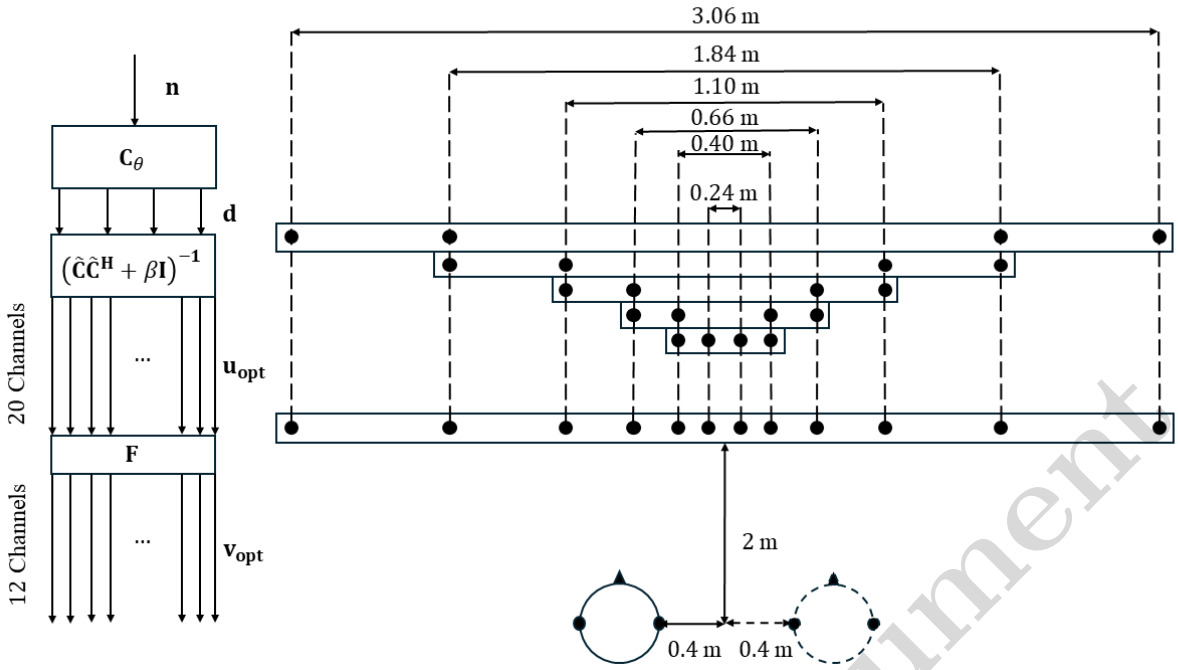

Figure 1: Array geometry (not to scale) and signal processing employed to generate array inputs from target signals.

which arises for each source azimuth, computed from a measured HRTF for each source azimuth. In the current work, azimuth predictions were calculated for each of 23 frequency channels, spaced in intervals of one equivalent rectangular bandwidth (ERB) between 200 Hz and 5000 Hz. For the lowest 12 frequency channels (up to a centre frequency of 1296 Hz), the azimuth was based on the fine-structure ITD. For the higher 11 frequency channels the azimuth estimate was based on the envelope ITD. For the envelope estimate, the ITD was mapped to azimuth based on the lowest frequency band in the lookup table. For each sample, these were pooled into a fine-structure and envelope estimate by mean averaging. To facilitate more direct comparison with the May model, these sample-wise azimuth estimates were further pooled into 56 10 ms time frames.

3. Results and Discussion

Figures 2-4 show histograms of the localization estimates of the two models, and the listening test data, for binaural stimuli generated only with the dummy head in the target position. Results are shown for target azimuths progressing (in reading order) from 0° in 30° increments anti-clockwise. The models always predict source positions in the front, and so listening test results are always corrected to the front. Clearly, the May model produces lower variance than is seen in the listener responses. For each target azimuth, the May model predicts responses close to the target even when this was not the case for listener responses, such as for a target azimuth of −90°. For the fine-structure Dietz predictions, the variance in model predictions also appears lower than the listener responses in most cases. Unlike the May model, the fine-structure Dietz model does not predict accurate localization of the most lateral target azimuths. The range of azimuth responses predicted by the envelope Dietz model are further compressed, with responses limited to between around −50° and 50°. A possible explanation for this is the high frequency content of the target stimuli. Because the Dietz model relies mostly on the ITD cue, it is likely unable to establish reliable azimuth estimates over much of the frequency range of the stimuli. Figure 5 shows the standard deviation of the listener responses, Dietz envelope, Dietz fine-structure and May model predictions. All the models under-predict the standard deviation in listener responses. In particular, the fine-structure Dietz model substantially under-predicts standard deviation for negative azimuths (left side) and the May model under-predicts standard deviation for positive azimuths (right side).

The standard deviation of the model estimates increases when variation in listener position is introduced to the model inputs. Figures 6-8 show the localization estimates produced by the May and Dietz models for 100 binaural signals. Each binaural signal is computed based on an HRTF for x and y co-ordinates drawn from normal distributions with mean 0 and standard deviation 4 cm and 6 cm, respectively. Each selected listener position was rounded to the closest 2 cm, the resolution with which HRTFs were measured. These distributions were selected so that model listeners would be positioned

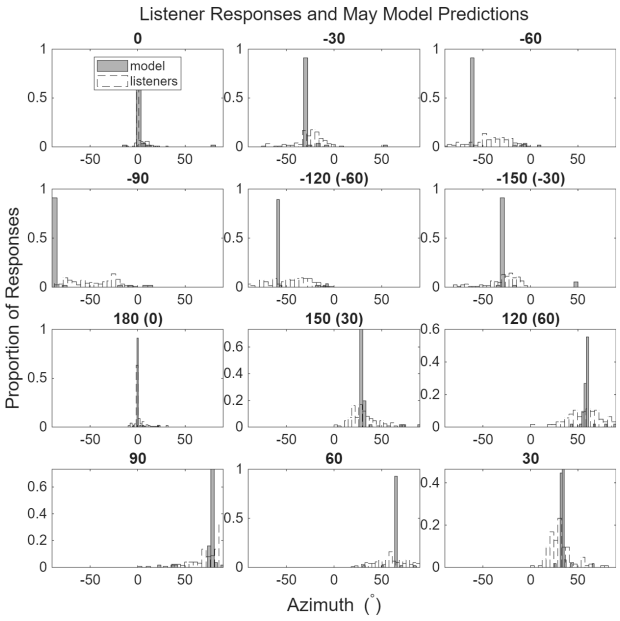

Figure 2: Distribution of listener responses and predictions by May model for a listener located in the target position. Target azimuths progress in 30° increments anti-clockwise in a reading order

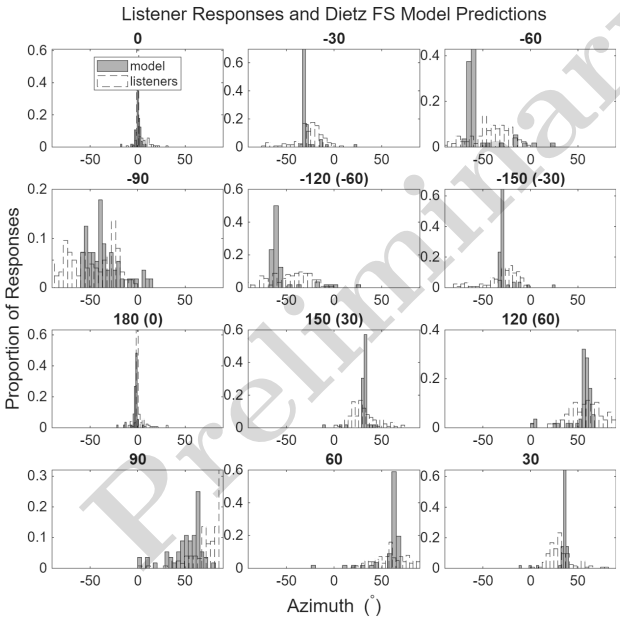

Figure 3: Distribution of listener responses and predictions by Dietz model for a listener located in the target position. Predictions are based on the fine-structure cues in the lowest 12 frequency channels (236 Hz - 1296 Hz). Target azimuths progress in 30° increments anti-clockwise in a reading order

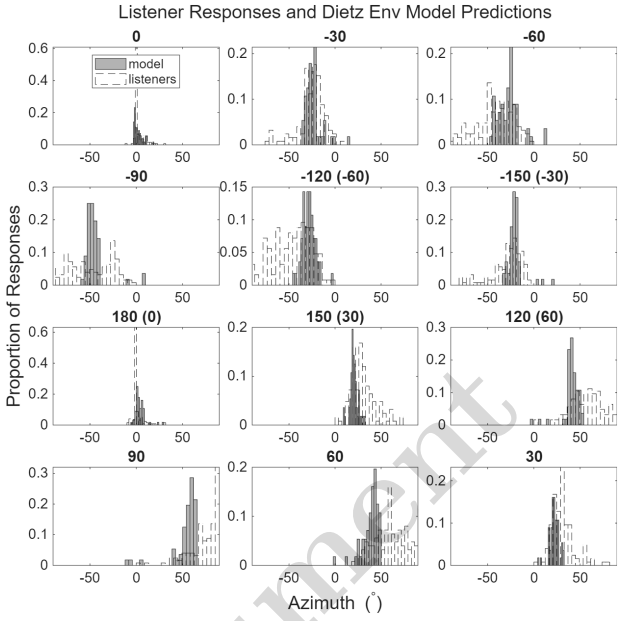

Figure 4: Distribution of listener responses and predictions by Dietz model for a listener located in the target position. Predictions are based on the envelope cues in the highest 11 frequency channels (1470 Hz - 4770 Hz). Target azimuths progress in 30° increments anti-clockwise in a reading order

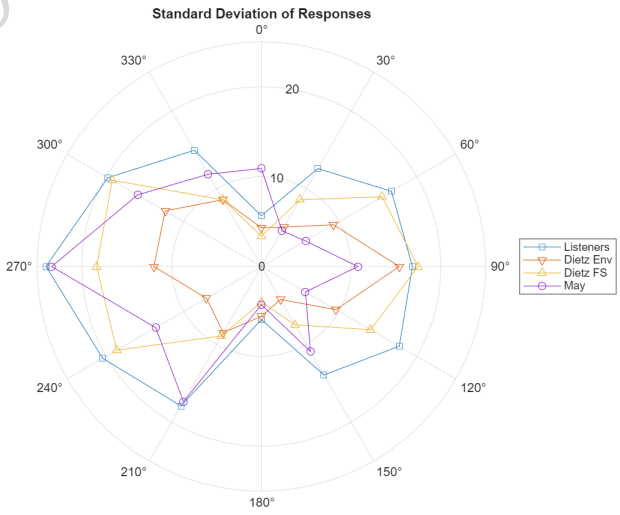

Figure 5: Standard deviation (in degrees) of listener responses and predicted responses by May model and Dietz model using fine-structure and envelope cues. Predictions are based on binaural inputs for a listener located in the target position. Fine-structure cues are based on lowest 12 frequency channels (236 Hz - 1296 Hz). Envelope cues are based on highest 11 frequency channels (1470 Hz - 4770 Hz).

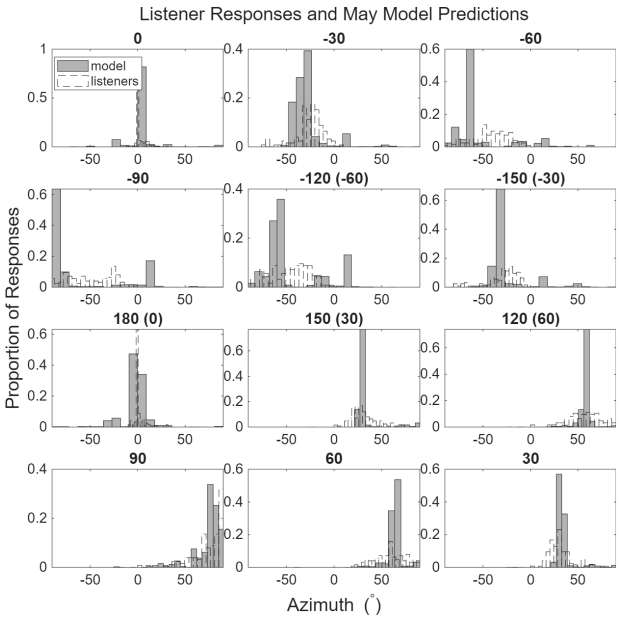

Figure 6: Distribution of listener responses and predictions by May model for normally distributed listener positions around the target position. Target azimuths progress in 30° increments anti-clockwise in a reading order

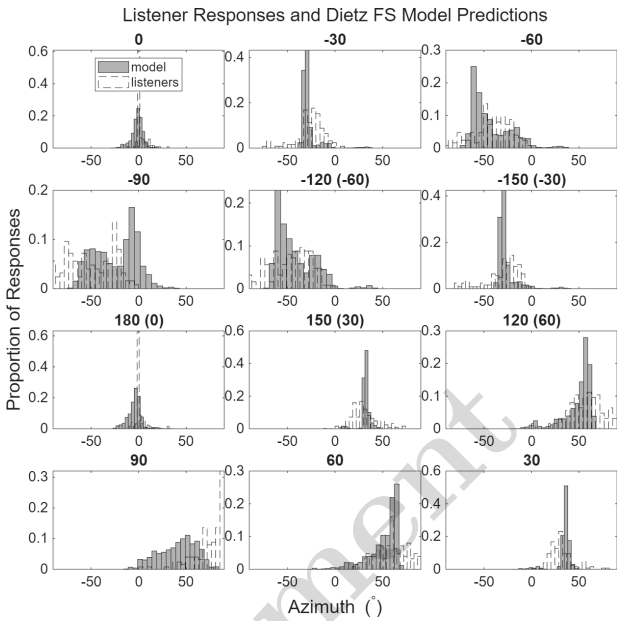

Figure 7: Distribution of listener responses and predictions by Dietz model for normally distributed listener positions around the target position. Predictions are based on the fine-structure cues in the lowest 12 frequency channels (236 Hz - 1296 Hz). Target azimuths progress in 30° increments anti-clockwise in a reading order

within a 10 cm by 10 cm square surrounding the target position for around 95% of model inputs. The variance in the May predictions has clearly increased here, although less difference is immediately apparent in the Dietz models. Again, the Dietz models fail to predict accurate localization for the most lateral target azimuths (−90° and 90°). Figure 9 shows the standard deviation of the listener responses, Dietz envelope, Dietz fine-structure and May model predictions for the same distribution of listener positions. In terms of standard deviation, the May model over-predicts for negative azimuths (left side). Both models predict the trend in listener responses of larger variance for target azimuths on the left side than on the right.

4. Summary

Existing binaural localization models based on ITD and ILD are capable of predicting listener responses to a novel two-listener virtual imaging system. Both the models tested were capable of predicting broad trends in the standard deviation seen in listener responses, such as greater variance in localization estimates for target azimuths to the left. Furthermore, as with the listener responses, the models predicted greater deviation in listener responses for the most lateral target azimuths. In general, the Dietz model under-predicted the azimuth for the most lateral target source angles, which was not seen to the same extent in listener responses. The May model, however, over predicted standard deviation for target azimuths to the left and under predicted standard deviation for target azimuths to the right.

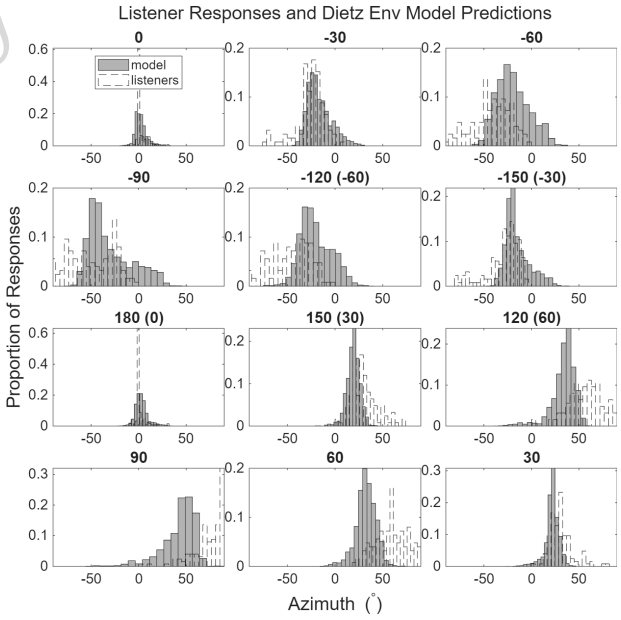

Figure 8: Distribution of listener responses and predictions by Dietz model for normally distributed listener positions around the target position. Predictions are based on the envelope cues in the highest 11 frequency channels (1470 Hz - 4770 Hz). Target azimuths progress in 30° increments anti-clockwise in a reading order

Figure 9: Standard deviation (in degrees) of listener responses and predicted responses by May model and Dietz model using fine-structure and envelope cues. Predictions are based on binaural inputs for normally distributed listener positions around the target position. Fine-structure cues are based on lowest 12 frequency channels (236 Hz - 1296 Hz). Envelope cues are based on highest 11 frequency channels (1470 Hz - 4770 Hz).

5. Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC,UKRI) EP/X032558/1

References

[1] L. A. Jeffress, "A place theory of sound localization," *Journal of Comparative and Physiological Psychology*, vol. 41, no. 1, pp. 35–39, 1948. DOI: 10.1037/h0061495.

[2] E. C. Cherry and B. M. A. Sayers, ""human 'cross-correlator'"—a technique for measuring certain parameters of speech perception," *The Journal of the Acoustical Society of America*, vol. 28, no. 5, pp. 889–895, 1956. DOI: 10.1121/1.1908506. [Online]. Available: http://dx.doi.org/10.1121/1.1908506.

[3] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. i. simulation of lateralization for stationary signals," *The Journal of the Acoustical Society of America*, vol. 80, no. 6, pp. 1608–22, 1986.

[4] W. Gaik, "Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling," *The Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 98–110, 1993.

[5] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Communication*, vol. 53, no. 5, pp. 592–605, 2011. DOI: 10.1016/j.specom.2010.05.006. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2010.05.006.

[6] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, 2011. DOI: 10.1109/TASL.2010.2042128. [Online]. Available: http://dx.doi.org/10.1109/TASL.2010.2042128.

[7] G. McLachlan, P. Majdak, J. Reijniers, and H. Peremans, "Towards modelling active sound localisation based on bayesian inference in a static environment," *Acta Acustica*, vol. 5, 2021. DOI: 10.1051/aacus/2021039. [Online]. Available: http://dx.xoi.org/10.1051/aacus/2021039.

[8] G. McLachlan, P. Majdak, J. Reijniers, M. Mihocic, and H. Peremans, "Dynamic spectral cues do not affect human sound localization during small head movements," *Frontiers in Neuroscience*, vol. 17, 2023. DOI: 10.3389/fnins.2023.1027827. [Online]. Available: http://dx.doi.org/10.3389/fnins.2023.1027827.

[9] J. Reijniers, G. McLachlan, B. Partoens, and H. Peremans, "Ideal-observer model of human sound localization of sources with unknown spectrum," *Scientific Reports*, vol. 15, no. 1, p. 7289, 2025, ISSN: 2045-2322.

[10] R. Barumerli, P. Majdak, M. Geronazzo, D. Meijer, F. Avanzini, and R. Baumgartner, "A bayesian model for human directional localization of broadband static sound sources," *Acta Acustica*, vol. 7, p. 12, 2023, ISSN: 2681-4617.

[11] H. Wierstorf, A. Raake, and S. Spors, "Binaural assessment of multichannel reproduction," in *The Technology of Binaural Listening*, J. Blauert, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 255–278, ISBN: 978-3-642-37762-4. DOI: 10.1007/978-3-642-37762-4_10. [Online]. Available: https://doi.org/10.1007/978-3-642-37762-4_10.

[12] Majdak, Piotr, Hollomey, Clara, and Baumgartner, Robert, "Amt 1.x: A toolbox for reproducible research in auditory modeling," *Acta Acust.*, vol. 6, p. 19, 2022. DOI: 10.1051/aacus/2022011. [Online]. Available: https://doi.org/10.1051/aacus/2022011.

[13] V. S. Paul, I. J. Lambert, and P. A. Nelson, "Localization performance of a two-listener virtual sound imaging system," Manuscript under review at Journal of the Acoustical Society of America, 2026.

[14] B. Bernschütz, *Spherical far-field hrir compilation of the neumann ku100*, Jul. 2020. DOI: 10.5281/zenodo.3928297. [Online]. Available: https://doi.org/10.5281/zenodo.3928297.