PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Auditory models comparison for horizontal localization of concurrent speakers in adverse acoustic scenarios

Roberto BARUMERLI[(1)], Andrea ALMENARI[(1)], Michele GERONAZZO[(2)], Giorgio Maria DI NUNZIO [(1)], and Federico AVANZINI[(3)]

[(1)]Dept. of Information Engineering, University of Padova, Italy, [barumerli,almenari,dinunzio]@dei.unipd.it
[(2)]Dept. of Architecture, Design, and Media Technology, Aalborg University, Denmark, mge@create.aau.dk
[(3)]Dept. of Computer Science, University of Milano, Italy, federico.avanzini@di.unimi.it

Abstract

This paper aims at comparing and reproducing the predictions of two public available computational auditory models for speaker localization in different simulated environments. The direction-of-arrival (DOA) of sound sources in the horizontal plane can be extracted by using binaural spatial cues from room and user acoustics. Since our predictions consider the specificity of both models at the level of peripheral processing, the proposed solution for DOA extraction also provides a common multi-conditional training for the Gaussian Mixture Model (GMM) approach. A set of acoustic simulations of adverse conditions (i.e. multi speakers or high reverberant scenarios) supports the evaluation phase on robustness of the synthetic auditory process. Our analysis reproduces two case studies from the scientific literature in order to investigate the reliability of localization predictions in the frontal horizontal plane. Finally, a newly defined acoustic scenario allows to identify differences between auditory models outcome in the entire horizontal plane. The results show a good agreement with previous literature and our machine learning approach emphasizes peculiarities of each approach for auditory peripheral processing.

Keywords: auditory model, binaural processing, room acoustic simulation

1 INTRODUCTION

Humans continuously perceive the environment using a multi-modal approach. Among other senses, hearing allows to locate objects coming from any direction and distance. The ability to understand the direction-of-arrival (DOA) of a sound source relies on the monaural and binaural acquisition of the acoustic environment [2]. While the anatomy of the external ears gives the ability to estimate the source elevation [1], the distance between the two pinnas allows the auditory cortex to evaluate the direction of a sound source on the horizontal plane [2]. Such elaboration also enables the separation of a specific source from a noisy environment [12]. This process offers the possibility to develop computational auditory models inspired on this biological mechanism [5], [18], [16]. The widely used features used to exploit the DOA by the human auditory cortex are the interaural time difference (ITD) and the interaural level difference (ILD) [2]. Usually, on top of these features, a machine learning approach provides an estimation of the sound source direction even more precise than the human counterpart [15]. Such models find their applications in different research fields: from cochlear implants to robot audition [14]. While a generic approach is suitable for a robotic system, using these methods can be interesting to automatically profile a single person from an acoustic and physiologically point of view [7, 10] for future audio technologies able to adapt to listeners[21]. Since the underlying biological process is not yet fully understood (i.e. [24]) different models rely on distinct assumptions leading to singular implementations. The attempt of this work was to compare two models, proposed in [18] and [5], which rely on the place theory [13] and on the rate code [3], respectively. Both model's implementations are publicly available into the Auditory Modelling Toolbox [22] and they estimate the DOA of single speech source in different acoustic environments for the frontal horizontal hemi-field. Such comparison aims to understand if different integrations and assumptions on the auditory pathway components can yield to similar results in term of estimation's precision. The proposed

evaluation is divided into two steps. First, we reproduce two experiments proposed in the scientific literature, one per model, to understand the reliability of the open-source implementations. Second, we force the same machine learning approach on both models aiming to extend the evaluation of DOA for the entire horizontal plane for two different acoustic scenarios. The paper is structured as follow: Section 2 proposes an overview of the models and the tools adopted for the simulations. Section 3 describes the implemented experiments and, finally, Section 4 reports and discusses the obtained results.

2 MATERIALS

2.1 May's auditory model

The auditory model implemented by May *et al.* in [18] relies on two features, ITD and ILD, for the DOA estimation limited to the frontal horizontal plane at zero elevation. The model adopts the common used place theory developed by Jeffress [13], and it estimates the DOA relying on a machine learning model. The auditory front-end consists of a Gammatone filter-bank followed by inner hair cell modelling. Each channel of the binaural sound is decomposed with a 4th-order Gammatone filterbank of 32 elements [15]. The center frequencies of the filter-bank are equally distributed between 80 Hz and 5 kHz by means of equivalent rectangular bandwidths (ERBs). Phase compensation for each channel is introduced to synchronize the binaural cues over the same time window. Moreover, the gain of each filter is adjusted to follow to the middle-ear transfer function. Then, a half-wave rectification with square-root compression is used to simulate neural transduction. Binaural cues are calculated using a $20ms$ window, with overlap of $10ms$, with a sampling frequency $f_s = 44.1kHz$. Each auditory channel returns the ILD as the ratio of the energy integrated in the time window between left and right ears, and the ITD is extracted for every frame as $(\tau + \delta)/f_s$ where τ is the time lag which maximizes the normalized cross-correlation and δ is the peak position relative to τ. To improve the ITD estimation an exponential integration is applied around τ. Hence, for each window the feature space is represented in Eq. 1 where index i represents the Gammatone filter and T the number of observations.

$$X_i = \{x_{i,1}, ..., x_{i,T}\} = \{(itd_i(1), ild_i(1)), ..., (itd_i(T), ild_i(T))\} \tag{1}$$

Finally, azimuth estimation is achieved through a machine learning approach based on the gaussian mixture model (GMM). For each Gammatone filter a GMM is trained to derive the azimuth α_i from the integration of the features composed of ITD and ILD pairs computed for that specific channel. The training of each GMM is performed by means of the iterative Expectation-Maximization (EM) algorithm by considering different source-receiver directions on different positions relative to a simulated reverberated room.

2.2 Dietz auditory model

The second model for azimuth estimation was proposed by Dietz *et al.* in [5] and it extends a previous work of the same authors [6]. The model is able to estimate a sound source direction on the horizontal plane at zero elevation by relying the ITD and ILD features. In order to compute the ITD this model introduced a variation to the Jeffress model: the contralateral inhibition. Such diversification allowed to determine the ITD by computing the differences on the auditory nerve's firing rates. This approach seems to be more tight to the human auditory processing [3] from a functional point of view. Moreover this model is based on a higher number of physiological findings than the May's one. The auditory front-end computed monoaural parameters by implementing the following processing pipeline: the middle ear is modelled with a 500-2000 Hz first-order band-pass filter; a 4th-order Gammatone bandpass filter-bank of 23 elements spaced of 1 ERB in 200-5000 Hz range which represented the basilar membrane; a cochlea compression is applied with an instantaneous compression with power 0.4; neural transduction of inner hair-cells with half-wave rectification and a 770 Hz fifth-order low-pass filter; finally, temporal disparities were extracted with a second-order complex Gammatone filter [6]. After this processing, complex signals $g_l(t)$ and $g_r(t)$ for both ears were obtained, each one with amplitude $a(t)$ and phase $\phi(t)$. The output of the peripheral processing is split by a fine-structure filter and modulation filter. This choice introduced a better temporal resolution, since fine-structure information is important for frequencies

lower than $1.4kHz$, and modulation becomes considerable above that frequency. In each separate process, the interaural transfer function ($ITF(t)$) is calculated:

$$ITF(t) = g_l(t) \cdot \bar{g}_r(t) = a_l(t) \cdot a_r(t) \cdot e^{j(\phi_r(t) - \phi_l(t))} \tag{2}$$

and the interaural phase difference (IPD) is derived as follow $IPD(t) = arg([ITF(t)]_{lp})$. ITD is extracted by dividing IPD by the mean of instantaneous frequency of left and right signals. To derive ILD, a second-order modulation low-pass filter with 30 Hz cut-off frequency is used both for left and right signals, obtaining an energy ratio between right and left signal in dB. Moreover, a mask on IPD is introduced in order to extract reliable segments: the interaural vector strength (IVS). This method captured the IPD fluctuation and it is used as an alternative of interaural coherence (IC) which cannot be calculated because this model does not rely on cross-correlation. In order to estimate the DOA, a ninth-order polynomial fit returned $\alpha_1 = p_f(\mid IPD \mid)$ and $\alpha_2 = p_f(\mid 2\pi - IPD \mid)$ and the sign of ILD resolved the ambiguity. Finally, a histogram is computed aggregating each azimuth estimations for all the frames considered and, in order to estimate the DOA, a sum of seven Gaussian functions fitted the histogram. The main peak returned the speakers' direction. [1]

It is worthwhile to mention that the HRTF dataset adopted in the original article are not publicly available consequently the replication section relied on the KEMAR mannequin dataset [8] since the model did not require a specific training.

2.3 Training the models

For the second part of this work, we extended the models to estimate the DOA for the full horizontal plane. The training method was also based on the GMM, but we adopted the multi conditional training proposed by Ma *et al.* [16]. The training set consisted in synthetic binaural samples generated by the anechoic MIT HRTF dataset limited to the horizontal plane with zero elevation [8] convolved with the TIMIT speech corpus [9] and combined with diffuse noise at different signal-to-noise ratios (SNRs). The diffuse noise was generated by the sum of 72 uncorrelated, white Gaussian noise sources, each one assigned to every HRTF direction. A set of 30 sentences was randomly selected for each azimuth value. Finally, the anechoic signal was corrupted with the diffuse noise at three SNRs {0, 10, 20} dB and, for both models, ITD and ILD were computed. For the Dietz's model the unwrapped ITD was used. Such approach allowed to train the models without any knowledge about the acoustic conditions of the evaluations. The training was performed with the EM algorithm available in the NetLab toolbox [19], which was also adopted for the May's model. Due to the elevated training time of Dietz model, the EM algorithm was limited to 100 iterations.

A model selection phase on the number of centers was carried out in order to minimize the training error. Due to elevated computational resources and simulation time, the number of samples was reduced, for each step, to three samples for training and eight samples for testing. The chosen values were eleven centers for May's model, with a training error of 14.12%, and nine centers for Dietz's model, with a training error of 26.70%. An error was accounted if the absolute difference between the estimated and true direction exceeded 5°.

3 SIMULATIONS

The first two experiments stated in both [5] and [5] manuscripts have been reproduced in order to evaluate the implementation publicly available in the AMToolbox [17]. Each experiment performance was evaluated for both models. Furthermore, in order to compare the different assumptions of the models, a new training phase guaranteed the comparison of the models' signal processing techniques. The entire set of experiments used the the TIMIT corpus [9], resampled to 44.1 kHz and normalized to 0 dB, and the HRTF dataset measured on the KEMAR mannequin [8] wrapped into a SOFA file [17].

[1] The histogram fitting by the sum of Gaussian functions was not implemented in the current version of the Auditory Modelling Toolbox. To add this feature, we implemented this step with the Curve Fitting Toolbox available in the MATLAB software.

Frequency [Hz]	125	250	500	1000	2000	4000
α	0.096	0.118	0.176	0.253	0.325	0.406

Table 1. Absorption coefficients adapted to obtain RT_{60}=0.69s.

3.1 Experiment's replicas

3.1.1 Concurrent speakers in a free-field environment

The experiment reproduced the analysis proposed in the Sec. 4.3 of the original paper [5]. The output of this simulation allowed to understand the scattering on the azimuth estimation with low SNR values. The conditions considered a scenario composed by three different speakers immersed in a speech shaped noise. The sources were positioned respectively on $\alpha = 30°$, $\alpha = 0°$ and $\alpha = -30°$ and the background noise scattered over all directions accounted $\alpha \in \{-80°, -30°, 0°, 30°, 80°\}$. The SNR was computed on the average power of the speaker at $\alpha = 30°$ to the total noise power. The source-receiver distance was imposed at 3 m and the SNRs were $0, -6$ dB.

3.1.2 Reverberated environment with different source/receiver configurations

This evaluation reproduced the setup proposed into the Sec. *V.C* of the May *et al.* original work [18]. The authors' experiment estimated model's accuracy in a reverberated room for different source-receiver combinations. The original computation relied on the acoustic simulator developed by Campbell *et al.* [4]. Since this toolbox was not available, the binaural room impulse responses (BRIR) were synthesized with the software proposed by Schimmel *et al.* in [20], which included the computation of the diffuse surface reflections leading to a more realistic rendering. For the simulation, the room had sizes of $5.1 \times 7.1 \times 3m$ with an averaged reverberated time (RT_{60}) of $0.69s$ (based on Sabine formula). This simulation did not account the diffusion computation, since Campbell's software did not provide it, and an adaptation of the reflection coefficients was necessary to obtain the original reverberation time (see Table 1). The source-receiver combinations considered eight different receiver positions in the room with 21 different source with directions from $-50°$ to $50°$, spaced by $5°$, and with distance of 1.5 m. Under these conditions, the evaluation procedure has been carried out for every receiver and every source position with four different cases: from one to four random sources with a minimum spacing of $10°$ between every concurrent source. The experiment estimated anomalies in azimuth estimation, where an anomaly was accounted if the estimated azimuth differed at least of $5°$ from the true value.

3.2 The new experiments

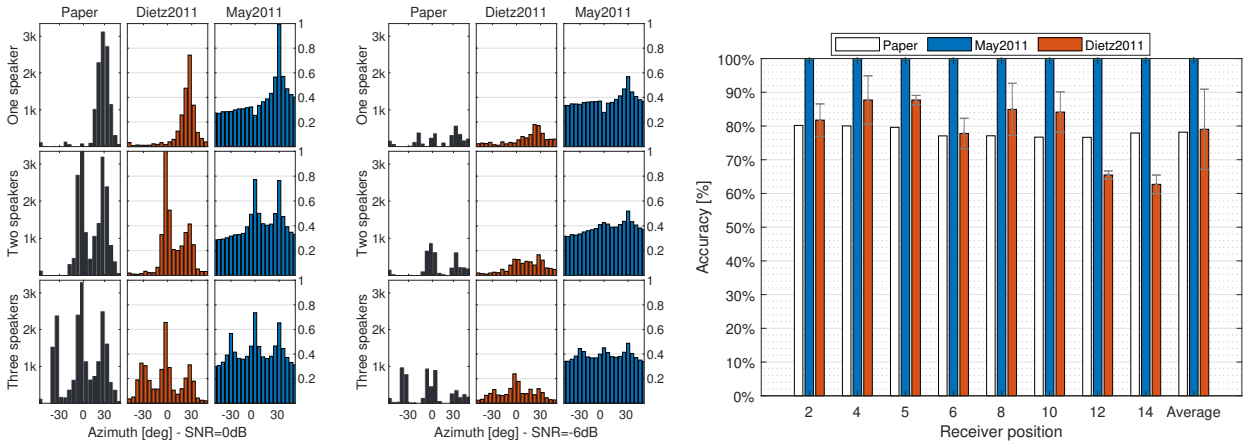
We extended both models to estimate the DOA for the entire horizontal plane at zero elevation. The HRTF dataset grid was defined in the interval $[0°, 355°]$ with a $5°$ spacing. These evaluations aimed to explore and measure also the front-back confusion. The experiments shared similar conditions to the experiment in Sec. 3.1.2: for each receiver and for each azimuth a number of one, two or three concurrent speakers was used at a fixed distance of 1.4 m randomly spaced at least of $10°$. All the samples had fixed length of 2.3 s and the evaluations were repeated five times each one with different speech samples. The adopted metric accounted a positive estimation if the predicted azimuth did not differ more than $5°$ from one of the concurrent speaker. Moreover, the guessed angle was corrected if the front-back confusion occurred by mirroring the estimation across the inter-aural axis.

3.2.1 Reliability on different reverberating conditions

This experiment explored the models reliability on the combination source-receiver positions with room reverberation time's variation, which was achieved by imposing different materials for the room's surfaces. Three rooms were simulated with fixed dimensions $5.1 \times 7.1 \times 3$ m and fixed receivers' positions {A=(1.5, 1.5, 1.75), B=(3.55, 1.5, 1.75), C=(4.05, 3.05, 1.75)}m. Table 2 reports the sets of absorption and diffusion coefficients considered. These values were derived from different materials measured in a real environment [23].

Frequency [Hz]	125	250	500	1000	2000	4000	Diffusion	Room1	Room2	Room3
Ceramic Tiles	0.01	0.01	0.01	0.02	0.02	0.02	0.3			floor
Wood panel	0.15	0.10	0.06	0.08	0.10	0.05	0.1		floor	walls
Carpet on concrete	0.02	0.06	0.14	0.37	0.60	0.65	0.1	walls	walls	
Seating people	0.55	0.86	0.83	0.87	0.90	0.87	0.5	floor		
							RT_{60}	0.37s	0.98s	1.88s

Table 2. Absorption coefficients and room configurations deployed for the experiment. RT_{60} was computed with the Sabine formula.



(a) (b)

Figure 1. Experiment replicas. Figure 1a: Concurrent speakers in a free-field environment. The x-axis indicates the azimuth grid while the y-axis the number of estimations accounted for a specific azimuth. Since each model used a different time window analysis the histograms were normalized in order to obtain a comparable shape. Figure 1b: Source-receiver combination experiment for different position in the simulated room. The errors' standard deviation over three repetitions were also reported.

3.2.2 Cocktail party simulation

This setup simulated a free field scenario were one ore more concurrent speech samples were corrupted by a diffuse speech shaped noise. The noise was generated by imposing for each direction 100 TIMIT sentences with randomized starting positions. Three different SNRs, {0, 10, 20} dB, allowed a comparison on the model's reliability against noise. The SNR was computed as the mean power of the sum from one to three speech samples over the sum of the diffuse noise.

4 RESULTS AND DISCUSSION

4.1 Experiment's replicas

The replica of the experiment from the work of Dietz *et al.* showed a good agreement with the original results (see Fig. 1a). The speech sources' directions were represented by the peaks in the histograms. In the original, plots the peaks were more prominent also for the lowest SNR value: this difference can be attributed to the different the HRTF dataset and samples employed for our simulations. Interestingly, Dietz's model with the IVS removed yielded to non reliable estimations by decreasing the overall height of the histogram's bins.
In Fig. 1b, results obtained on the evaluation of source-receiver in the reverberant room are reported. Here, the models performed mostly better than the May *et al.* original work [18]. Two factors could have led to this

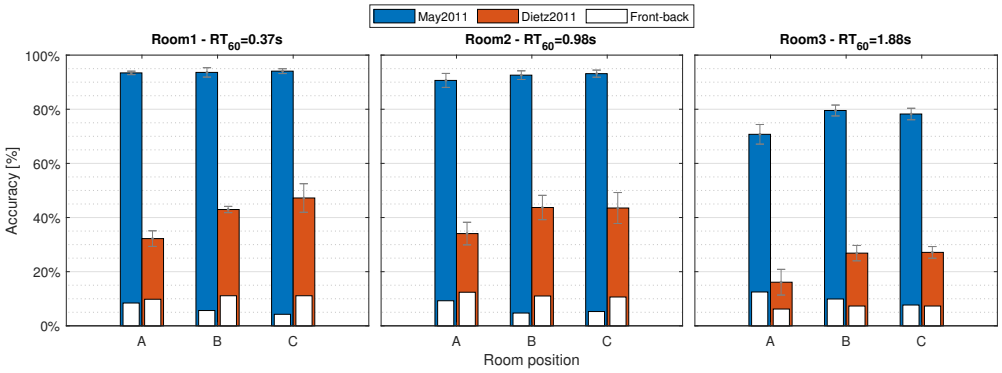


Figure 2. Simulation' results for different reverberation times and receiver positions. The front-back confusion was resolved before the accuracy computation.

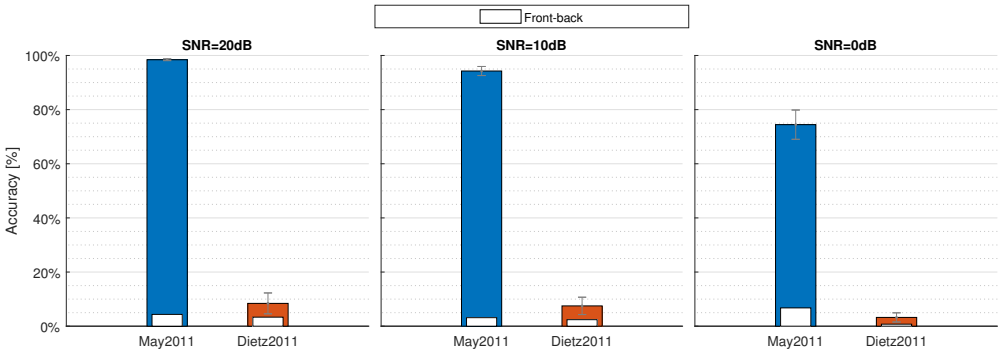


Figure 3. Simulation of localization estimation of speech sample immerse in a diffuse speech shaped noise. The front-back confusion was resolved before the accuracy computation.

results: the different room simulator, with different RIR's rendering implementation, or the lack of information about the method used to estimate the RT_{60}. In this work, we adopted the Sabine formula which accounted the room size and the absorption coefficients.

4.2 New experiments

The results of the simulations with different RT_{60} is reported in Fig. 2. The May's model resulted in a better accuracy with respect to the Dietz's. While the front-back confusion remained below the 15% for all conditions, the accuracy decreased with the increase of the reverberant time. The trials in position A demonstrated an overall worse accuracy and front-back confusion: this was probably related to the symmetrical distance from the walls leading to similar reverberant paths coming from orthogonal directions. In Fig. 3, the free field simulation with a speech shaped diffuse noise is reported. The main trend showed that both model's averaged accuracy decreased with the increment of the noise power in respect of the speech sources. The lack of precision of the Dietz's model was probably related to the computation techniques of the binaural features. An example of the differences between the two models is reported in Fig. 4. The May's model was reliable in the extended scenario while the Dietz's model showed a similar trend but with very different accuracy. The May's model also showed a similar tendency with the psychoacoustic literature [11]. The obtained differences were probably due to the lack of coordination between the Gammatone filterbank: the May's model adjusted gain and phase of each filter in order to compensate the group delay of each channel. This approach led to a better generalization of the feature space than one implemented in the Dietz's model. Moreover the extended processing introduced

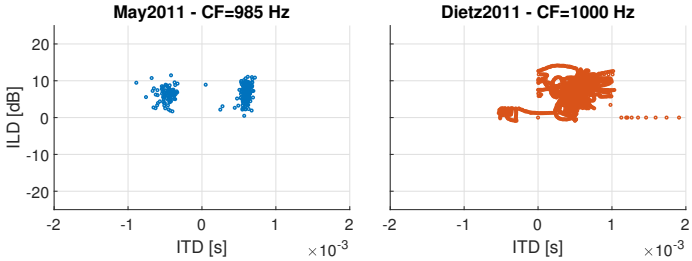


Figure 4. Computed binaural features for the nearest Gammatone filter with Characteristic Frequency, CF, near f=1kHz, for a speech source at $\alpha = 60°$.

by Dietz *et al.* was intended for a tight simulation of the human auditory process lacking a reliability and qualitative analysis of the feature space, which was limited for an HRTF dataset of five directions in frontal horizontal plane.

5 CONCLUSIONS

In this work, the experimental results of two different auditory models were reproduced and extended. The reproduction encountered some disparities in the simulations due to the lack of information in the original works but the results showed a good agreement. The new simulations showed that the Dietz's model needs an improvement from the implementation point of view in order to understand if the accuracy can be increased. This model adopted the rate code for the ITD computation that probably did not receive the same broad adoption as the place theory, which is implemented in the May's model, where the cross correlation finds an easier integration in the signal processing approach. Despite the encountered differences, it is believed that an effort to increase the fidelity of an auditory model can lead also to an improvement of its precision. This achievement can extend reliability of such system that could outperform the human accuracy but also lead towards a better profile of a single listener from an acoustic point of view. Further research is needed in order to improve the outcome of similar comparisons for a better understanding of which auditory model is the best choice for profiling human listening abilities.

REFERENCES

[1] R. Barumerli, M. Geronazzo, and F. Avanzini. Localization in Elevation with Non-Individual Head-Related Transfer Functions: Comparing Predictions of Two Auditory Models. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2539–2543, Sept. 2018.

[2] J. Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[3] J. Breebaart, S. van de Par, and A. Kohlrausch. Binaural processing model based on contralateral inhibition. I. Model structure. *The Journal of the Acoustical Society of America*, 110(2):1074–1088, Aug. 2001.

[4] D. Campbell, K. Palomaki, and G. Brown. A MATLAB simulation of" shoebox" room acoustics for use in research and teaching. *Computing and Information Systems*, 9(3):48, 2005.

[5] M. Dietz, S. D. Ewert, and V. Hohmann. Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication*, 53(5):592–605, May 2011.

[6] M. Dietz, S. D. Ewert, V. Hohmann, and B. Kollmeier. Coding of temporally fluctuating interaural timing disparities in a binaural processing model based on phase differences. *Brain Research*, 1220:234–245, July 2008.

[7] M. Dietz, J.-H. Lestang, P. Majdak, R. M. Stern, T. Marquardt, S. D. Ewert, W. M. Hartmann, and D. F. M. Goodman. A framework for testing and comparing binaural models. *Hearing Research*, 360:92–106, Mar. 2018.

[8] W. G. Gardner and K. D. Martin. HRTF measurements of a KEMAR. *The Journal of the Acoustical Society of America*, 97(6):3907–3908, June 1995.

[9] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, 93, 1993.

[10] M. Geronazzo, E. Peruch, F. Prandoni, and F. Avanzini. Applying a single-notch metric to image-guided head-related transfer function selection for improved vertical localization. *J. Audio Eng. Soc.,*, 67(6):1–15, June 2019.

[11] M. L. Hawley, R. Y. Litovsky, and H. S. Colburn. Speech intelligibility and localization in a multi-source environment. *The Journal of the Acoustical Society of America*, 105(6):3436–3448, June 1999.

[12] S. Haykin and Z. Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005.

[13] L. A. Jeffress. A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, 41(1):35–39, 1948.

[14] Jens Blauert. *The technology of binaural listening*. Modern Acoustics and Signal Processing. Springer, 2013.

[15] Lyon, Richard F. *Human and machine hearing*. Cambridge University Press, 2017.

[16] N. Ma, T. May, and G. J. Brown. Exploiting Deep Neural Networks and Head Movements for Robust Binaural Localization of Multiple Sources in Reverberant Environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2444–2453, Dec. 2017.

[17] P. Majdak, Y. Iwaya, T. Carpentier, R. Nicol, M. Parmentier, A. Roginska, Y. Suzuki, K. Watanabe, H. Wierstorf, H. Ziegelwanger, and others. Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions. In *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

[18] 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*, 19(1):1–13, Jan. 2011.

[19] I. Nabney. *NETLAB: Algorithms for Pattern Recognition*. Springer Science & Business Media, 2002.

[20] S. M. Schimmel, M. F. Muller, and N. Dillier. A fast and accurate "shoebox" room acoustics simulator. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 241–244, Apr. 2009.

[21] S. Serafin, M. Geronazzo, N. C. Nilsson, C. Erkut, and R. Nordahl. Sonic Interactions in Virtual Reality: State of the Art, Current Challenges and Future Directions. *IEEE Computer Graphics and Applications*, 38(2):31–43, 2018.

[22] P. L. Søndergaard and P. Majdak. The Auditory Modeling Toolbox. In *The Technology of Binaural Listening*, pages 33–56. Springer, Berlin, Heidelberg, 2013.

[23] Z. Sü and S. Yilmazer. The Acoustical Characteristics of the Kocatepe Mosque in Ankara, Turkey. *Architectural Science Review*, 51(1):21–30, Mar. 2008.

[24] E. Verschooten, S. Shamma, A. J. Oxenham, B. C. Moore, P. X. Joris, M. G. Heinz, and C. J. Plack. The upper frequency limit for the use of phase locking to code temporal fine structure in humans: A compilation of viewpoints. *Hearing Research*, 377:109–121, June 2019.