APPLICATION OF AN AUDITORY PROCESS MODEL FOR THE EVALUATION OF STEREOPHONIC IMAGES

M. Park,
P.A. Nelson
and F. Fazi

Institute of Sound and Vibration Research, University of Southampton,
Highfield, Southampton SO17 1BJ, UK

K.O. Kang

Electronics and Telecommunications Research Institute (ETRI),
138 Gajeongno, Yuseong-gu, Daejeon, 305-700, Republic of Korea

1 INTRODUCTION

Human ability to localise sound sources has long been studied, and the results of some classical experiments are summarised by Blauert [1]. In recent studies, the target area has been fully extended to three dimensions, and there has been great improvement in the test equipment, including the electromagnetic head-tracking device that has enormously facilitated data acquisition with a high level of accuracy (e.g. see Makous and Middlebrooks [2] and Carlile et al. [3]).

As the relevant technology to provide virtual sound fields advances from the classical system of stereophony to recent multi-channel surround systems, subjective evaluation of phantom images has been also of great research interest. As quoted by Rumsey [4], listening tests in the early days were mostly carried out to investigate the localisation of virtual images created by the conventional two-channel stereophony system. In more recent studies, the fidelity of virtual images has been tested for the lateral configuration of loudspeakers, where the evaluation and the optimisation of quadrophony and 5.1 channel surround systems were the primary objectives [5-7].

The experimental study to be presented in this paper is similar to the previous work summarised above, where the subjective responses to acoustic images will be investigated in terms of spatial accuracy. The baseline measurement of localisation accuracy will first be established by using a monopole sound source in the horizontal plane. In the following listening tests, a relatively wide range of stereophonic arrangements will be investigated, including symmetric and asymmetric loudspeaker locations to a listener's front, side and back. These results obtained in the subjective tests will then be compared to the predictions of a hearing model that has been recently suggested by the authors [8-10]. The hearing model considers the excitation-inhibition (EI) cell activity patterns [11] (EI patterns hereinafter) as an internal representation of sound localisation cues, for which the central process finds the best match from the EI-pattern template predefined for free-field stimuli in the horizontal plane.

The arrangement of the listening tests will be detailed in section 2, followed by a brief description of the hearing model in section 3. The two results from the subjective tests and the model simulation will then be analysed and compared in section 4, and finally some conclusions will be given in section 5.

2 LISTENING TESTS

2.1 Method

A total of 10 university personnel (7 male and 3 female) have participated in the listening tests. Pure tone audiometry showed that all participants have acceptable hearing ability across the frequency range of interest (0.5~8 kHz, <20 dB).

The current listening tests have been carried out in a small anechoic chamber at the University of Southampton, which approximately measures $5m \times 5m \times 3m$. This chamber is annexed by a control room where most of the equipment was placed, and the experimenter had a CCTV facility to monitor the subject inside the room.

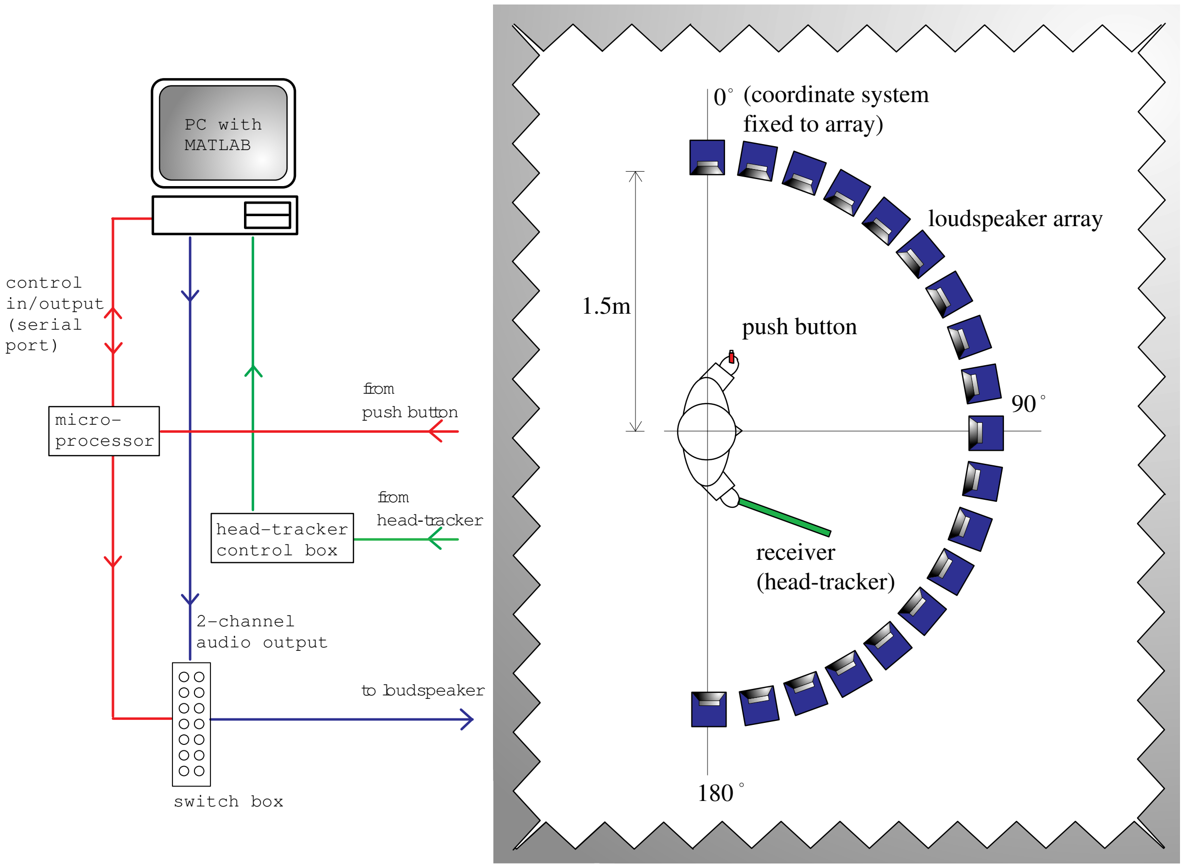

Figure 2.1 Diagram illustrating the test and the control rooms.

An array of 19 loudspeakers was located in the test room at every 10° from 0° to 180° with respect to the room coordinate system (see Figure 2.1). The height of the array was approximately adjusted to the average sitting height of the participants, while the distance between the loudspeakers to the centre of the array was measured to be 1.5 *m*. Since the visual cues given by the loudspeakers may bias the subjective judgements of acoustic image positions, the array has been covered by thin black curtains with rigid metal wires placed on top of each loudspeaker unit, which was extended beyond the loudspeakers at both ends by ~20 *cm*. In this way, absolute and relative positions of loudspeakers could not be recognised by the participants.

100-*ms* white Gaussian noise with 6-*ms* rise-fall ramps has been used as a stimulus. For the real source localisation tests, this signal has been presented to the listener through only one of the loudspeakers. On the other hand, for the localisation of stereophonic images, amplitude gains, g_1 and g_2 have been applied to the stimulus to produce the 2-channel input signals to a chosen pair of loudspeakers. Target image position, θ_t, and the positions of the two active loudspeakers, θ_1 and θ_2 are related to the amplitude gains, g_1 and g_2 by the following equations [7].

$$\theta_m = \frac{\pi}{2} \times \frac{\theta_t - \theta_1}{\theta_2 - \theta_1}, \quad (\theta_1 \le \theta_t \le \theta_2) \tag{2.1}$$

$$\begin{aligned} g_1 &= \cos\theta_m \\ g_2 &= \sin\theta_m \end{aligned} \tag{2.2}$$

This panning scheme is intended to keep the total acoustic power constant. When a single loudspeaker was used with a unit gain, the sound pressure level at the centre of the array was calibrated to be 70 dBA.

Subjects participated in 5 one-hour sessions, and each session was divided into 5~6 blocks that each contained 20 trials. A single trial started with a recorded message which instructed the listener to direct his/her head to the initial frontal position marked with a cross. When ready, the listener pressed a push button to play the stimulus, and reported the image position by using a pointing device equipped with an electromagnetic position tracker (Polhemus FASTRAK) and a laser pointer. On hearing the stimulus, the listener held this pointer to the perceived image position, and switched on the laser pointer for visual confirmation. Then pressing the push button once again saved the tracker reading to the computer, and the next trial started.

2.2 Image positions under investigation

The following four categories of acoustic image positions have been tested in the current experiment.

Category 1 – Localisation of single sound source: While only a single loudspeaker was being used throughout the two repeated sessions, a total of 10 subjective judgements (5 in each session) have been obtained for target locations from 0° (listener's front) to 180° at every 10°. Test results in this category can be regarded as the individual baseline performance of the sound localisation task.

Category 2 – Stereophony systems with 60° angular aperture: Standard stereophonic arrangement with 60° angular aperture was tested at the listener's front, side and back. Given the loudspeaker configuration, stereophonic images were created at every 10° between transducers.

Category 3 – Various lateral arrangements: In a pilot experiment, it has been found that it is hard to create convincing virtual images at listener's side. For this reason, the influence of the angular aperture has been investigated in a series of tests in category 3, where the position of the second loudspeaker, θ_2 was fixed at 90° (listener's right), while θ_1 varied from 30° to 70°.

Category 4 – 5.1 channel surround system: In the conventional 5.1 channel configurations (ITU-R BS.775), loudspeakers are located at 0° (C), 30° (R), 110° (RS), 250° (LS) and 330° (L). In the last category of the current test, 3 adjacent loudspeaker pairs of the standard 5.1 channel configuration (C-R, R-RS and RS-LS) have been tested, where image position varied at every 10° between the chosen transducers).

While the loudspeaker array designed for the current listening test covers only a half circle, the above listed target positions required loudspeaker configurations to the subject's front, side and back. Accordingly, the initial sitting direction had to be adjusted for some of the target positions, and, to minimise seat relocations, the above 4 categories of target positions have been regrouped into 6 sessions. In each session, all target conditions have been randomly tested for 5 times for each participant. Therefore, a total of 50 subjective responses could be obtained for each virtual image position presented by a specific stereophony setup (100 responses for each real source position).

3 HEARING MODEL

The binaural hearing model based on the EI pattern-matching technique has been described in detail elsewhere (e.g. see Park et al. [8]). Given the binaural signals at the ear drums as the model input, the signal transformation at peripheral, binaural and central stages are simulated by corresponding signal processing modules, and the final output of the model is the prediction of acoustic image position. In the following paragraphs, the simulations using the current model will be briefly described.

3.1 HRTF measurement

In order to establish hearing models, head-related transfer functions (HRTF) have been first measured for each participant of the listening tests. The measurement was carried out in a large anechoic chamber, where a single loudspeaker was mounted at the same height as the subject's ears. After an initial positioning procedure, HRTFs have been recorded by using small in-ear microphones, while the subject's chair was rotating 5° at a time. The position of the subject was monitored by a head-tracking device (Polhemus FASTRAK), and an automated voice-feedback system operated to guide the subject to maintain the initial position.

3.2 Peripheral processes

According to the loudspeaker configurations and the target positions described in section 2.2, binaural signals can be synthesised by convolving the stimulus with the measured individual HRTFs [12]. These synthesized signals are the input to the peripheral processes of the model, where various signal processing modules simulate the transfer characteristics of the middle and the inner ears. In particular, the frequency selectivity of the basilar membrane has been modelled by a fourth-order gammatone filterbank with 60 channels from 300 Hz to 12 kHz. Compared to similar hearing models, the density of this filterbank appears to be reasonable [13], one half of each equivalent rectangular bandwidth (ERB) [14] overlapping with the nearby filters.

The 60 channel intermediate signals are further processed to take into account the transduction of neural impulses in the organ of Corti. First, a half-wave rectifier removes the negative part of the signals, and then the gradual loss of phase-locking is simulated by a low-pass filter cut-off at 770 Hz [11]. Finally, the nonlinear relation between the input and the output level is approximated by a square-root compressor [15].

3.3 Binaural processes

The equalisation-cancellation (EC) processor suggested by Breebaart et al. [11] generates EI patterns in each auditory frequency band, combining the neural impulse signals from peripheral processors on the left and right channels. The EI patterns are the curved surface over the domain of the characteristic interaural time difference (ITD) and interaural level difference (ILD), and the minimum of the pattern indicates the most probable ITD and ILD associated with the input binaural signals. In addition, it has been shown that the EI patterns are nearly unique with respect to the azimuthal position of a sound source and the frequency [10].

Given the EI patterns, the imperfect hearing processes, for example, the untimely transduction of neural impulses and the noise from internal organs, can be approximated by a Gaussian noise mask added to the EI patterns. In this way, the model prediction can now be considered to be a random variable with certain mean and variance values.

3.4 Central processes

In the central process, the EI patterns are compared to the collection of EI patterns (forming the template) that have been generated for each monopole source located between 0° and 359° at every 1°. This matching procedure based on the cross-correlation between the EI patterns will produce 60 'local' estimates in each auditory frequency band, which may then be frequency-weighted to give a probability density function associated with the perceived image position. The power spectral density of the intermediate signals from the peripheral processor combined with the weighting function suggested by Raatgever (see Stern et al. [16]) has been employed as a tentative frequency-weighting scheme.

4 TEST RESULTS COMPARED TO MODEL PREDICTIONS

Localisation responses in the listening tests have been pooled across subjects, and means and variances have been obtained. On the other hand, 100 responses have been sampled in the simulation from each of the ten individual models for a specific acoustic image position (it should be recalled that the EI noise mask is a random variable), and the model responses have been also pooled across individual models, and averaged.

<u>Category 1 – Localisation of single sound source</u>

The result of the real source localisation is first presented in Figure 4.1. The two results from the test and the simulation appear to be similar at least in terms of the local features observed near 90°. However, it is apparent that the subjective data are significantly lower than the model predictions, especially in the mid-range target position, and the difference can be as much as 10°. This underestimation of the image position is considered to have resulted from the measurement bias, perhaps related to the mispositioning of the participants. For example, the relatively recent data from Carlile et al. [3] show far less mean errors, agreeing better with the current simulation results. In particular, without any head-restraint or chin-resting facility in the current listening test, the participant could move off the initial centre position, although they were instructed not to do so. Especially, they tend to make forward-backward movement more easily than in the lateral direction, which could have incurred greater measurement errors at the side, whereas the frontal and the rear sources have been localised relatively accurately. However, without any accompanying test results, it is not clear what caused this seemingly systematic bias, which should certainly be taken into account in analysing the test data presented in the remaining sections.

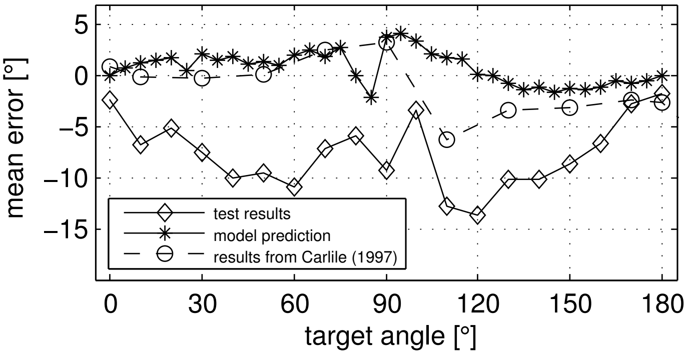

Figure 4.1 Mean errors in the localisation of single sound source

<u>Category 2 – Stereophony systems with 60° angular aperture</u>

Figure 4.2 shows the standard frontal stereophony system with 60° angular distance between loudspeakers. Test results show that the virtual images are perceived at the location which monotonically increases with the amplitude ratio. However, the image position is greatly underestimated, which, as discussed in the case of real source localisation, reflects some bias that may have resulted from the subject mispositioning. A similar observation could be made for the stereophony system to the back of the

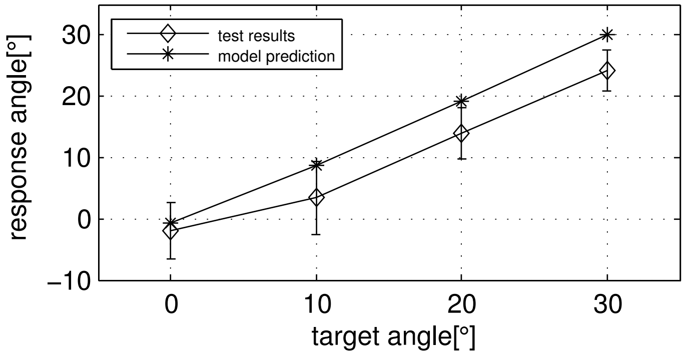

Figure 4.2 Localisation of stereophonic images presented by standard frontal stereophony setup

listener [loudspeakers at 210° (L) and 150° (R)], where the variance of subjective responses was greater compared to the frontal stereophony (data not shown).

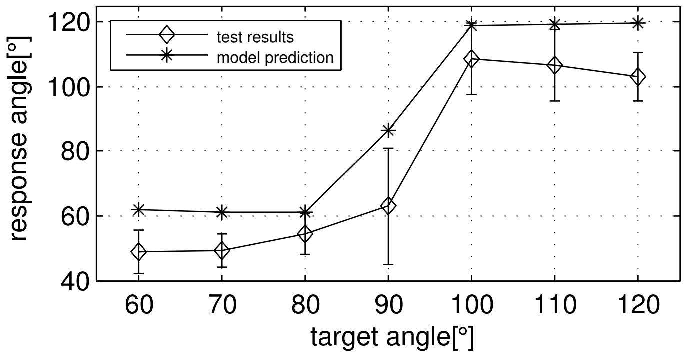

Figure 4.3 Localisation of stereophonic images presented by front-back symmetric loudspeaker arrangment with 60DEG angular aperture

The results for the lateral stereophony system are shown in Figure 4.3, which has two loudspeakers symmetrically positioned with respect to the frontal plane. It is interesting to see that the image position is not monotonically related to the relative amplitude ratio. In other words, subjective responses are mainly found at the position of the louder transducer, and shift rather abruptly around 90°. The stereophony system based on amplitude-panning has been designed to exploit the path length difference between transducer and receiver positions. Therefore, it is not surprising to see that the lateral setup is ineffective in controlling the stereophonic image position, for the path lengths are identical from each ear to loudspeakers. The model predicts it very well, and it makes prediction even without committing front-back confusion, already resolved possibly by the subtle difference of ITD-ILD cue pairs for the front and back.

<u>Category 3 – Various lateral arrangements</u>

From the results presented in category 2, it is apparent that the virtual acoustic scene to the side of listener may not be efficiently created by a front-back symmetric loudspeaker configuration. Therefore, in the following test, one of the loudspeakers has been considered to be fixed at 90°, while the position of the other varied between 30° and 70°. The result in Figure 4.4 shows that the position of stereophonic image is not linearly related to the target position. Instead, the acoustic image rather slowly moves from the first loudspeaker position, and then at a certain point, at approximately 80°, it makes a swift movement to the second loudspeaker position. The deviation from the designated image position is greater with a wider loudspeaker angular aperture. Features found for this category of test have been successfully predicted by the simulation results of the model (see Figure 4.5), although the subjective bias that underestimates the image position is still noticeable.

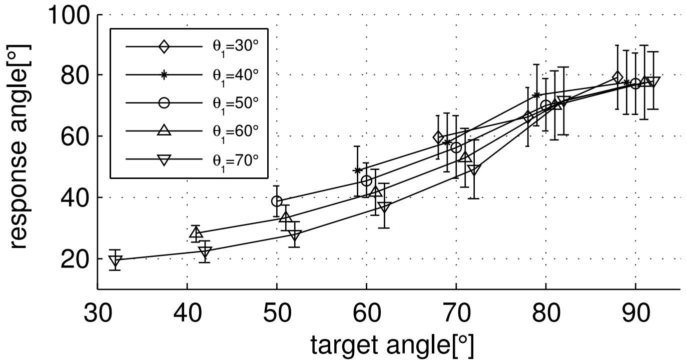

Figure 4.4 Localisation of stereophonic images presented by various loudspeaker arrangements to the side of listener – Subjective responses

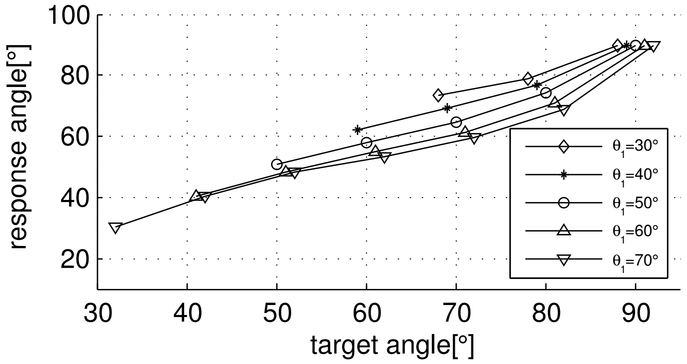

Figure 4.5 Localisation of stereophonic images presented by various loudspeaker arrangements to the side of listener – Model prediction

Category 4 – 5.1 channel surround system

Finally, test and simulation results for the 5.1 channel system are presented in Figure 4.6. In the conventional configuration, acoustic images to the side of the listener are presented by the two lateral loudspeakers at 30° (330°) and 110° (250°). In the listening test, it has been shown that this transducer pair asymmetric to the frontal plane is capable of better controlling the acoustic image position than the symmetric arrangement investigated in category 2. Nevertheless, it is also noteworthy that the virtual images can be still ambiguous near ±90°, particularly between 100° (240°) and 120° (260°). The image ambiguity at the lateral positions is predicted well by the model, and, in general, the simulation results agree reasonably well with the subjective responses.

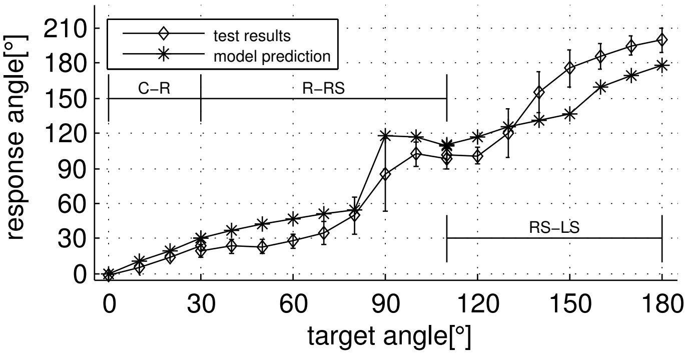

Figure 4.6 Localisation of stereophonic images presented by conventional 5.1 channel surround system

5 CONCLUSION

In this study, various stereophonic loudspeaker arrangements have been investigated in the listening tests. Target acoustic image positions have been found to be often underestimated up to 10°, which may have resulted from some systematic bias associated with the test arrangement and protocol. Nevertheless, some important features of the stereophony system have been identified, particularly regarding the perceptual quality of acoustic images presented to the listeners' side, which depends on the absolute and relative position of the loudspeakers. In the accompanying simulation study, these features have been reasonably well predicted by the hearing model which has been established from the individual HRTF databases. Given the successful application to the conventional stereophony and 5.1 channel surround systems, the current hearing model is expected to be useful for the perceptual evaluation of spatial audio systems in general, and also the output of the associated signal processing algorithms and coding techniques.

6 ACKNOWLEDGEMENT

This work was supported by the IT R&D program of MIC/IITA. [2007-S004-01, Development of Glassless Single-User 3D Broadcasting Technologies]

7 REFERENCES

1. Blauert, J., *Spatial Hearing: The Psychophysics of Human Sound Localization*. 2001, London: MIT Press.
2. Makous, J.C. and J.C. Middlebrooks, *2-Dimensional Sound Localization by Human Listeners.* Journal of the Acoustical Society of America, 1990. **87**(5): p. 2188-2200.
3. Carlile, S., P. Leong, and S. Hyams, *The nature and distribution of errors in sound localization by human listeners.* Hearing Research, 1997. **114**(1-2): p. 179-196.
4. Rumsey, F., *Spatial Audio*. 2001, London: Focal Press.
5. Martin, G., et al., *Sound source localization in a five-channel surround sound reproduction system*, in *107th AES Convention*. 1999: New York.

6. Theile, G. and G. Plenge, *Localization of Lateral Phantom Sources.* Journal of the Audio Engineering Society, 1977. **25**(4): p. 196-200.
7. West, J., *Five-channel panning laws: an analytical and experimental comparison*. 1998, University of Miami.
8. Park, M., P.A. Nelson, and Y. Kim. *An auditory process model for sound localization*. in *IEEE WASPAA*. 2005. New Paltz, New York.
9. Park, M., P.A. Nelson, and Y. Kim, *An auditory process model for the evaluation of virtual acoustic imaging systems*, in *120th AES convention*. 2006: Paris, France.
10. Park, M., P.A. Nelson, and Y. Kim. *An auditory process model for the evaluation of virtual acoustic imaging systems*. in *IOA*. 2006. Southampton, U.K.
11. Breebaart, J., S. van de Par, and A. Kohlrausch, *Binaural processing model based on contralateral inhibition. I. Model structure.* Journal of the Acoustical Society of America, 2001. **110**(2): p. 1074-1088.
12. Moller, H., *Fundamentals of Binaural Technology.* Applied Acoustics, 1992. **36**(3-4): p. 171-218.
13. Jin, C., M. Schenkel, and S. Carlile, *Neural system identification model of human sound localization.* Journal of the Acoustical Society of America, 2000. **108**(3): p. 1215-1235.
14. Moore, B.C.J., *An introduction to the psychology of hearing*. 5 ed. 2003: Academic Press.
15. Patterson, R.D., M.H. Allerhand, and C. Giguere, *Time-Domain Modeling of Peripheral Auditory Processing - a Modular Architecture and a Software Platform.* Journal of the Acoustical Society of America, 1995. **98**(4): p. 1890-1894.
16. Stern, R.M., A.S. Zeiberg, and C. Trahiotis, *Lateralization of Complex Binaural Stimuli - a Weighted-Image Model.* Journal of the Acoustical Society of America, 1988. **84**(1): p. 156-165.