A binaural auditory model and applications to spatial sound evaluation

Marko Takanen[1], Gaëtan Lorho[2], and Matti Karjalainen[1]

1 Helsinki University of Technology, Dept. of Signal Processing and Acoustics, Espoo, Finland

2 Nokia Corporation, Helsinki, Finland

Reproduced sound quality is influenced by many perceptual factors such as sound source direction and distance, timbre, loudness, spatial impression, and temporal characteristics. Listening tests can be performed to assess these aspects, but computational auditory models provide an interesting alternative to the direct use of human subjects. While this model-based approach is commonly used for monaural audio evaluation, the application of advanced binaural models to spatial sound evaluation remains a challenge. In this paper, we present a binaural auditory model based on the two principles of interaural cross-correlation and binaural cue selection and we illustrate how it can be utilized in two different spatial sound applications.

INTRODUCTION

Spatial aspects play an important role in the perceived quality of reproduced sound. In stereophonic reproduction, the illusion of 3-D sound fields can be created using, e.g., amplitude panning, stereo enhancement algorithms or head-related transfer functions (HRTFs). Amplitude panning can be used to position sound events between the two loudspeakers and the positioning of these loudspeakers along with the applied panning determine the obtained spatial image. Stereo enhancement algorithms have also been developed to create virtual sources outside the loudspeaker span of stereophonic systems. HRTFs on the other hand can be used to position sound events around the listener in headphone reproduction or in loudspeaker reproduction when the principle of crosstalk cancellation formulated by Atal *et al.* (1966) is applied. They can be obtained by acoustic measurements on human or artificial heads or via computer simulations.

The performance of these different methods is usually assessed by listening tests but an alternative to the direct use of human subjects can also be considered with the help of binaural auditory models. We exploited this approach in our work with the aim of assessing the characteristics of reproduced sound through instrumental metrics. The binaural auditory model presented in this paper is based on the coincidence structure by Jeffress (1948) and on the binaural cue selection model by Faller and Merimaa (2004). We describe first the functionality of the model including the peripheral hearing model, the way the binaural processing model is used to estimate thebinaural cue values in different frequency bands and the mapping approach selected to evaluate the direction of sound source(s) from the estimated values based on a set of reference

Proceedings of ISAAR 2009: Binaural Processing and Spatial Hearing. 2nd International Symposium on Auditory and Audiological Research. August 2009, Elsinore, Denmark. Edited by J. M. Buchholz, T. Dau, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN: 87-990013-2-2. The Danavox Jubilee Foundation, 2010.

values. Then, we demonstrate how this model can be utilized to analyze different HRTFs with the help of binaural cue values and we report how the stereo image width evaluated from the direction estimates changes when the loudspeaker span in the reproduction is increased. We conclude the paper with a short discussion.

BINAURAL AUDITORY MODEL

The binaural auditory model developed in this work can be used to derive either the binaural cue reference values or the direction(s) of sound source(s) from the input signal, which can be either a binaural recording made with an artificial head or an HRTF-processed source signal.

The structure of the model is illustrated in Fig. 1. The model consists of four parts, each of them having a specific task. The two first parts focus on modeling the peripheral hearing and binaural processing in the human auditory system while the other parts aim at evaluation of reference binaural cue values and the direction(s) of sound source(s). In this section, we present the four parts of the binaural auditory model in separate subsections.

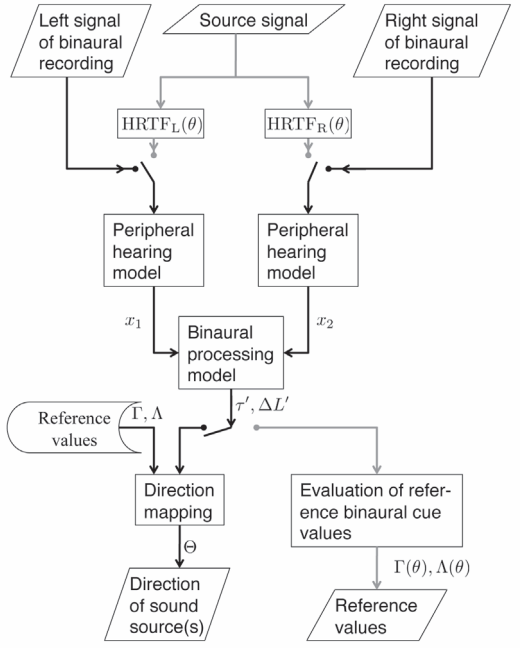

Fig. 1: Flowchart illustrating the structure of the developed binaural auditory model.

Peripheral hearing model

In the peripheral hearing model, the left and right ear signals are first passed through a middle ear compensation filter that resembles the transfer function presented by Moore *et al*. (1997) and are then processed with a cochlear model consisting of two steps. The first step in the cochlear model simulates the frequency selectivity of the basilar membrane with a 24-band gammatone filter bank (GTFB), whose center frequencies were set to match those of the critical bands on the Bark scale. This kind of filter bank structure was selected for computational efficiency reasons and we found this approach to give sufficient (spatial) resolution for the intended application(s) of the model. This GTFB was implemented with the freely available Matlab toolbox of Slaney (1998).

The second step models the functionality of the inner hair cells in the organ of Corti. The temporal function developed by Bernstein *et al*. (1999) was used in this step to simulate the pattern of the inner hair cell firings. This temporal model was implemented with the help of the freely available Matlab toolbox by Akeroyd (2001).

Binaural processing model

The binaural processing model uses the cross-correlation method to estimate the binaural cue values from the outputs of the peripheral hearing model. The physiological basis for this approach was proposed by Jeffress (1948). This model is based on the idea of binaural cue estimation presented by Faller and Merimaa (2004) and is implemented in three separate steps, which are 1) interaural cross-correlation (IACC) calculation and extraction of binaural cue values, 2) derivation of short-time loudness values and 3) binaural cue selection.

In the IACC step, the interaural time difference (ITD), interaural level difference (ILD) and interaural cohrence (IC) estimates are computed from the peripheral hearing model outputs following the binaural processor idea presented in Faller and Merimaa (2004). These estimates are computed separately for each frequency band. The obtained estimates for ITD, ILD and IC are denoted as τ, ΔL and c, respectively.

Derivation of short-time loudness values

The idea of (binaural) short-time loudness is exploited in the binaural cue selection process. In this model, we evaluate the short-time loudness values in non-overlapping time frames, the lengths of which are 20 ms. Consequently, the binaural short-time loudness values, denoted as $\hat{N}(z)$, are derived by

$$\hat{N}(z) = \sum_{e=1}^{2} \sqrt[4]{\frac{1}{I} \sum_{i=1}^{I} x_e^2 [z, i]}, \qquad \text{(Eq. 1)}$$

following the idea presented by Pulkki *et al*. (1999). Here the fourth root estimates the exponent presented by Zwicker and Fastl (1999), z denotes the frequency band, x denotes the peripheral hearing model output, I denotes the total number of samples in the time frame, and e denotes the left or right ear.

Binaural cue selection

The purpose of this step in the model is to select reliable localization cues. The number of criteria used in this step depends on the task the model is used for. If the model is employed to estimate the direction(s) of sound source(s) (see Fig. 1), a loudness criterion is used to select the reliable binaural cues. The value of this loudness criterion, denoted as $N_0(z)$, is computed separately for each frequency band based on the noise floor in the beginning of the binaural recording (i.e., during the first 60 ms) according to Eq. 1. The reliability of the binaural cue estimates is then evaluated by comparing the obtained loudness criteria and the short-time loudness values one time frame at a time. This comparison is made separately for each frequency band following the criterion, which states that the binaural short-time loudness on a given frequency band must exceed the loudness criterion value (i.e., $\hat{N}(z) > N_0(z)$), otherwise the binaural cue values in that frequency band are discarded.

If the model is employed to derive the reference binaural cue values (see Fig. 1), a criterion for the IC estimates is used additionally to the loudness criterion. This second criterion is used to ensure that the binaural cue reference values are accurate estimates of the free-field cues. This is accomplished by accepting the binaural cue estimates only when the coherence between the left and right ear signals is high (i.e., $c \approx 1$). In the current model, we used an IC criterion of 0:97 for the binaural cue selection.

Evaluation of reference binaural cue values

In this work, we selected a lookup table approach for the direction mapping process. The reference binaural cue values of this lookup table were derived from signals with known source direction[1]. In the process of obtaining these values, two parameters are computed separately for each frequency band from the selected binaural cue estimates, namely the sample mean and variance. These parameters are calculated separately for ITD and ILD estimates. Together with the known azimuth angle, denoted as θ, these parameters are used to represent the binaural cue reference values in the lookup table as

$$\Gamma[\theta, z] \sim N(\mu_{\mathrm{ITD}}(z), \sigma^2_{\mathrm{ITD}}(z)),$$
$$\Lambda[\theta, z] \sim N(\mu_{\mathrm{ILD}}(z), \sigma^2_{\mathrm{ILD}}(z)),$$

1 These signals can be generated for example by HRTF filtering at pre-defined azimuth angles.

where Γ and Λ denote the reference values for ITD and ILD, respectively, and μ and σ^2 denote the sample means and variances for the corresponding binaural cue values at different frequency bands.

Direction mapping

In the present model, we use a statistical approach to derive the most probable sound source direction(s) by comparing the selected binaural cue estimates to the reference values in the lookup table. This approach is based on two assumptions. Firstly, we assume that the signal in a given 20 ms time frame at a given frequency band can only correspond to a single direction. Secondly, we assume that the binaural cue estimates within a given time frame are randomly picked samples from one of the distributions in the binaural cue lookup table.

Taking into account these assumptions, the process of sound source direction mapping is applied separately for each time frame and comprises two phases. In the first phase, we test the hypothesis that the binaural cue estimates of a given time frame belong to the distribution of binaural cues associated with one (or several) angles of the lookup table. The hypotheses are tested separately for the ITD and ILD with the same 5% significance level. The information about the acceptance/rejection of these hypotheses are stored in variables h_{ITD} and h_{ILD}, the variables which can have values of either zero when the given hypothesis is accepted or one when the hypothesis is rejected. Also the probabilities behind the acceptance/rejection of the hypotheses are stored in variables p_{ITD} and p_{ILD}.

In the second phase, we use these four variables to derive the sound source directions separately for each frequency band. The most probable sound source direction, denoted as Θ, is selected from the group of accepted hypotheses based on the probabilities p_{ITD} and p_{ILD}, according to

$$\Theta[z,j] = \arg\max_{\theta} \begin{cases} p_{\mathrm{ITD}}\left[z,\theta\right], & z \leq 11 \wedge \exists\theta : h_{\mathrm{ITD}}(z,\theta) = 0, \\ p_{\mathrm{ILD}}\left[z,\theta\right], & z > 11 \wedge \exists\theta : h_{\mathrm{ILD}}(z,\theta) = 0, \end{cases} \qquad \text{(Eq. 2)}$$

where θ denotes the azimuth angle in the lookup table and j denotes the index of the inspected time frame. If none of the hypotheses for a given frequency band is accepted, the direction for that frequency band is replaced with a missing value. In this procedure, the direction mapping is divided into two separate frequency ranges based on the characteristic frequencies of the GTFB filters. For the frequencies below 1.5 kHz (i.e., the first 11 frequency bands) the direction mapping relies entirely on the ITD and above this frequency, only ILD is used to determine the direction of the sound source. This choice was motivated by the duplex theory of Lord Rayleigh stating that ITD and ILD are responsible for the localization in the low and high frequency ranges, respectively.

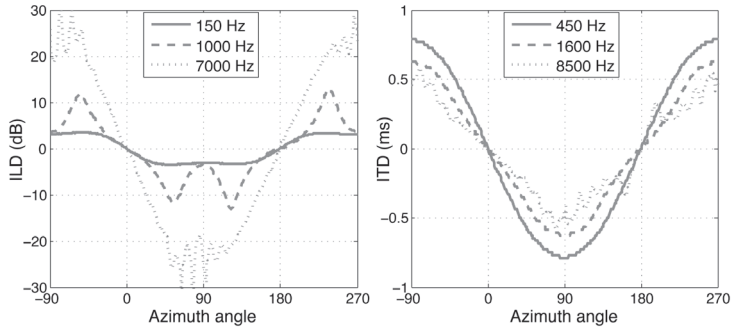

Fig. 2: Angle dependence of the binaural cue values evaluated from a set of HRTFs in the horizontal plane.

APPLICATIONS TO SPATIAL SOUND EVALUATION

In this section, we illustrate how the binaural auditory model developed in this work can be utilized in different spatial sound applications. An analysis of binaural cues in HRTFs will be presented first and a short study of the effect of stereophonic speaker span on stereo image width will then be reported.

Analysis of binaural cues in HRTFs

HRTFs can be analyzed with the auditory model to estimate binaural cues, which contain the information needed for spatial localization by the auditory system. In practice, a pink noise signal of short duration, e.g. 3 seconds, is filtered with a set of Head-related impulse responses (HRIRs) and this processed signal is passed through the model to produce a set of binaural cue estimates. This type of analysis can be applied to study the angle dependence of binaural cues in HRTFs and to assess the quality of HRTF filters as described next. To study the angle dependence of binaural cues in HRTFs, we selected a subset of responses covering the horizontal plane with a spatial resolution of one degree from the simulated HRTF database developed by Kirkeby *et al*. (1999). For each azimuth angle, an HRTF-processed signal was applied to the model and gave a constant ITD and ILD value for each of the 24 critical bands. We analyzed the variations of these values across azimuth angles by inspecting visually the values separately for each critical band. The two plots shown in Fig. 2 illustrate the pattern obtained for a subset of critical bands. This study illustrated that the binaural cues in this set of HRTFs vary smoothly across the horizontal plane in general but it also highlighted few interesting patterns. Firstly, the ITD pattern (right plot in Fig. 2) appears to be more regular across critical bands than the ILD pattern (left plot in Fig. 2). Secondly, the smoothness of the curves across angles depends on the critical band and higher frequencies are characterized by larger fluctuations, especially for the ITD. Thirdly, the ILD curves of critical bands in the mid range, e.g. 1 kHz in the left plot in Fig. 2, show a difference compared to the sinusoidal pattern around 90° and 270° azimuth, which corresponds to the bright spot at the contralateral side.

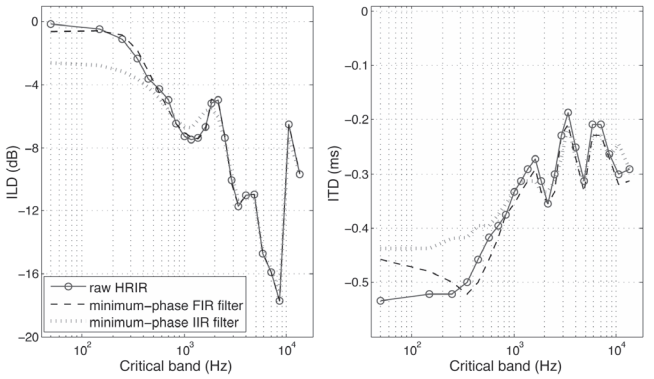

Fig. 3: Differences in the estimated binaural cue values between the original HRIRs and the two minimum-phase filters.

The application of binaural auditory models to HRTF filter quality assessment has been presented by Pulkki *et al*. (1999). We considered this approach in the present study by applying to our model a small set of HRIR filters from the simulated HRTF database developed by Kirkeby *et al*. (1999). The design of the binaural filters comprising an ipsilateral and a contralateral response followed the technique described by Huopaniemi and Smith (1999). We created first a minimum-phase approximation of the two original HRIRs from which we designed a 64-tap FIR filter by a simple windowing method and a 16-tap IIR filter by the Yule-Walker least-square design method. A frequency-independent ITD was also estimated by a method of cross-correlation between the two original HRIRs and this was approximated by a simple delay line inserted at the contralateral side of the the minimum-phase FIR and IIR binaural filters.

A pink noise signal was then filtered with the original set of HRIRs and the two HRIR approximations for several azimuth angles for binaural auditory model analysis. The resulting binaural cues were finally analyzed visually by plotting the ITD and ILD as a function of critical band for each binaural filter. Figure 3 illustrates the result for the 40° azimuth angle. The two minimumphase binaural filters show a relatively good fit to the original set of HRIRs overall but some differences can be seen in the lower critical bands for both the ITD and the ILD, especially for the IIR filter[2]. Also, the small shift seen between the reference and the two HRIR approximations in the ITD plot (right pane in Fig. 3) might come from the sample rounding introduced by the delay line employed for the frequency-independent ITD approximation.

In addition to the detailed information provided by these binaural cue plots, a global measure of error between the reference and the different target binaural filters can be defined from the ITD and ILD values giving therefore a perceptually-motivated quality estimation of HRTF filters.

2 It should be noted that techniques to emphasize the low-frequency in the filter design such as frequency warping (Huopaniemi and Smith, 1999) might have improved the IIR filter model fit but were not considered in this study.

Estimation of the stereo image width of loudspeaker setups

Another application of the binaural auditory model relates to the estimation of sound source direction in different loudspeaker setups. For this study, a set of binaural recordings was made with a Genelec 8020A loudspeaker setup and a Head and Torso Simulator (HATS B&K type 4128c) in non-anechoic near-field conditions (i.e. at 0.5 m distance in a listening room). To assess the effect of speaker angle on the stereo image width, we considered four different scenarios: 1) a monophonic reproduction at 0°, 2) a closely spaced stereophonic setup at ±10°, 3) an optimal stereophonic setup at ±30° and 4) a wide stereophonic setup at ±50°. We selected an instrumental jazz music track (*Home at last* by Steely Dan) to test the performance of the model with a complex signal. 10 seconds of these binaural recordings were analyzed by the model and the output of the direction mapping algorithm was analyzed visually. The results are presented in the form of a probability of sound source presence per azimuth angle for each scenario in Fig. 4. The effect of the loudspeaker span can clearly be seen between the very narrow sound source distribution in the monophonic case and the wider distribution of the stereophonic setups. The ±50° speaker setup shows the widest stereo width with sound events detected between -40° and -70°. The fact this range goes beyond the speaker span is due to an issue with the direction mapping algorithm that needs further investigation.

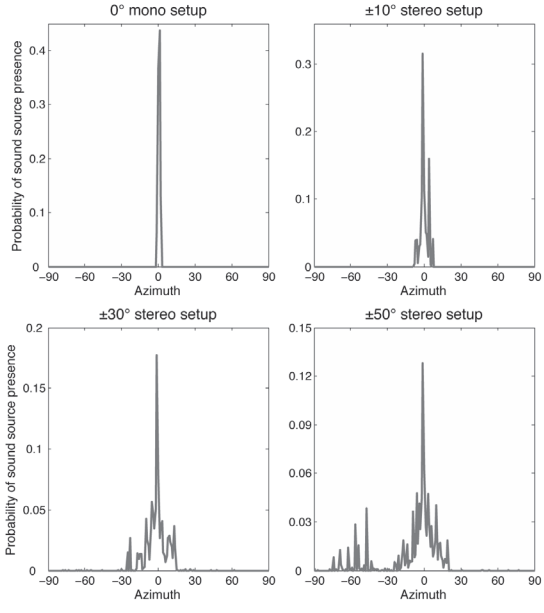

Fig. 4: Estimated stereo image width of the four different loudspeaker setups.

DISCUSSION

In this paper we described a binaural auditory model and showed two of its applications to the evaluation of spatial sound. The proposed model was constructed by refining and combining elements from the models presented by Moore *et al*. (1997), Slaney (1998), Bernstein *et al*. (1999), Faller and Merimaa (2004), and Pulkki *et al*. (1999). Implemented in Matlab, it takes binaural signals as input and evaluates either the binaural cue values or the direction(s) of sound source(s) from these signals. During this evaluation process the inputs, which can be either HRTF-processed source signals or binaural recordings, are passed through a series of processing blocks that simulate the human auditory system.

Previously, e.g., Huopaniemi *et al*. (1999) and Pulkki *et al*. (1999) have reported how HRTFs can be analyzed from the binaural cue values. In this paper we used a similar approach and showed how the angle dependence and smoothness of the binaural cues can be used to study HRTFs. We also demonstrated how the binaural cue values provided by the model can be exploited to compare different HRTF filter designs in terms of quality.

We also applied the model to evaluate the stereo image width from the direction estimates. This application illustrated that differences the loudspeaker span result in a clear change in stereo image width in frontal directions. At the same time this application showed also some issues with the direction mapping algorithm at the sides. These issues are related to a non-unique match to the reference values in the binaural cue lookup table. Therefore, one of the directions for future work is to investigate further these issues. Other possible direction for future work could be the estimation of the number of sound sources from the probability graphs taking into account aspects of auditory grouping in human auditory processing.

ACKNOWLEDGEMENTS

This work has been supported by the Academy of Finland (project no: 121252) and by Nokia corporation.

REFERENCES

Akeroyd, M. A. (**2001**). “A Binaural Cross-correlogram Toolbox for MATLAB,” http://www.ihr.mrc.ac.uk/products/index.php?page=matlab.

Atal, B. S., Hill, M., and Schroeder, M. R. (**1966**). “Apparent Sound Source Translator,” U.S. Patent 3236949.

Bernstein, L. R., van de Par, S., and Trahiotis, C. (**1999**). “”The normalized interaural correlation accounting for NoSπ thresholds with Gaussian and ”low noise” masking noise,” J. Acoust. Soc. Am. **106**, 870 – 876.

Faller, C., and Merimaa, J. (**2004**). “Source localization in complex listening situations: Selection of binaural cues based on interaural coherence,” J. Acoust. Soc. Am. **116**, 3075 – 3089.

Huopaniemi, J., and Smith, J. O. (**1999**). "Spectral and time-domain preprocessing and the choice of modeling error criteria for binaural digital filters," in *16th Intl. Convention of the Audio Eng. Soc*. (Rovaniemi, Finland, April 10-12, 1999). pp. 301–312.

Huopaniemi, J., Zacharov, N., and Karjalainen, M. (**1999**). "Objective and Subjective Evaluation of Head-Related Transfer Function Design," J. Audio Eng. Soc. **47**, 218 – 239.

Jeffress, L. A. (**1948**). "A place theory of sound localization," J. Comp. Physiol. Psychol. **61**, 468 – 486.

Kirkeby, O., Seppälä, E., Kärkkäinen, A., Kärkkäinen, L., and Huttunen, T. (**1999**). "Some Effects of The Torso on Head-Related Transfer Functions," in *122nd AES Convention* (Vienna, Austria, May 2007).

Moore, B. C. J., Glasberg, B. R., and Baer, T. (**1997**). "A Model for the Prediction of Thresholds, Loudness and Partial Loudness," J. Audio Eng. Soc. **45**, 224 – 240.

Pulkki, V., Karjalainen, M., and Huopaniemi, J. (**1999**). "Analyzing Virtual Sound Source Attributes Using a Binaural Auditory Model," J. Audio Eng. Soc. *47*, 203 – 217.

Slaney, M. (**1998**). "Auditory Toolbox: Version 2," Apple Technical Report #1998-010.

Zwicker, E., and Fastl, H. (**1999**). *Psychoacoustics, Facts and Models* (Springer-Verlag, Germany). Second Updated ed.