A binaural model to predict position and extension of spatial images created with standard sound recording techniques

Jonas Braasch

1CIRMMT, Faculty of Music, McGill University, Montreal

Correspondence should be addressed to Jonas Braasch (jb@music.mcgill.ca)

ABSTRACT
A binaural model was used to investigate different microphone techniques (Blumlein, ORTF, MS, spaced omni). In contrast to previous attempts, the model algorithm was not only designed to predict the position, but also the spatial extent of a reproduced spatial image. The architecture of the model was designed to optimally process binaural cue distributions with multiple peaks as often found in psychoacoustical data. The model also contains elements to simulate the precedence effect, which is required for analyzing spaced-microphone techniques, and is also useful when measuring the influence of the concert space on the recording.

1. INTRODUCTION
Several stereo microphone techniques exist to capture spatial information in sound recordings. Typically, the spatial information is encoded by orienting microphone axes in different directions (e.g. Blumlein technique), spacing both microphones apart (e.g. spaced-omni techniques) or both (e.g., ORTF technique). The design of 2-channel stereophonic microphone set-ups is typically chosen such that the spatial image is preserved when both recorded microphone channels are reproduced by two loudspeakers in a standard stereo configuration. Here, the loudspeakers are typically placed equidistant from the listener at angles of −30° and 30° (see Fig. 1).

The aim of this investigation was to apply a binaural model to evaluate the performance of different microphone techniques. To the author’s knowledge two binaural models have been previously applied to investigate the localization curve for classic microphone techniques [16], [20]. Localization curves describe the relationship between the azimuth of the original source position and the azimuth of the auditory event when listening to the reproduced signal [27]. Both models are based on the analysis of interaural time differences (ITDs) and interaural...
Fig. 1: Standard stereo loudspeaker set-up.

The binaural model that was developed for this investigation analyzes interaural time differences (ITDs) and interaural level differences (ILDs) by calculating interaural cross-correlation and simulating excitation/inhibition cells. Both ITDs and ILDs are determined frequency-band wise using a gammatone-filter bank with 36 bands, covering a frequency range from 20 Hz to 20 kHz. Afterwards, the ITDs and ILDs are remapped to azimuth positions. The frequency-dependent relationship between ITDs and azimuth angles, and ILDs and azimuth angles, was gained through analysis of HRTF measurements. Elements of contralateral inhibition were implemented to simulate the precedence effect for both the ITD and the ILD analysis. In the decision device of the model, the position of a sound source is estimated by weighting and combining the remapped binaural cues. A virtual environment and measured impulse responses were used to simulate the pathway from the sound source, via the microphone recording and the loudspeaker reproduction, to the listener’s ears.

In the next section, the basic acoustical principles of microphone techniques are described, which will be later analyzed by the binaural model.

2. CLASSIC MICROPHONE TECHNIQUES

In a typical recording situation, the transfer function between the sound source (e.g., a musical instrument treated as a one-dimensional signal in time \( x(t) \)) and a microphone is determined by the distance and the orientation between the microphone’s directivity pattern and the instrument. The distance determines the delay \( \tau \) between the radiated sound at the sound-source position and the microphone signal \( y(t) \):

\[
\tau(r) = \frac{r}{c_s},
\]
with the distance $r$ in meters and the speed of sound $c_s$. The latter can be approximated as 344 m/s at room temperature (20° C). According to the inverse-square law, the sound pressure radiated by a sound source will decrease by 6 dB with each doubling of the distance $r$:

$$p(r) = \frac{p_0 \cdot r_0}{r}, \quad (2)$$

with the sound pressure $p_0$ of the sound source at a reference distance $r_0$. In addition, it should be considered that the sensitivity of a microphone varies with the angle of incidence ($\alpha$) according to its directivity pattern. A directivity pattern can be written in a simple general form:

$$\Gamma(\alpha) = a + b \cos(\alpha). \quad (3)$$

Typically, the maximum sensitivity is normalized to one:

$$a + b = 1, \quad (4)$$

and the different available microphones can be classified using different factors for $a$ and $b$:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>omni-directional</td>
</tr>
<tr>
<td>0.75</td>
<td>0.25</td>
<td>sub-cardioid</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>cardioid (uni-directional)</td>
</tr>
<tr>
<td>0.25</td>
<td>0.75</td>
<td>hyper-cardioid</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>figure-8 (bi-directional)</td>
</tr>
</tbody>
</table>

The overall gain $g$ between the sound source (treated as a point source) and the microphone can be determined as follows:

$$g = g_d \cdot \Gamma(\alpha), \quad (5)$$

with the distance-dependent gain $g_d$. The transfer function between the sound source and the microphone signal is determined by two parameters only, the gain $g$ and the delay $\tau$, if the microphone directivity patterns are considered to be independent of frequency and frequency-dependent energy losses of the traveling wave are neglected. The relationship between the sound radiated from the source $x(t)$ and the microphone signal $y(t)$ is found to be:

$$y(t) = g \cdot x(t - \tau). \quad (6)$$

The earliest technique to control the spatial image using two microphones is the classic XY microphone technique introduced by Alan Blumlein in 1931 [3]. Here, two bidirectional microphones are arranged at an angle of 90° in the horizontal plane (Fig. 2a). Theoretically, both microphone diaphragms are at the same location in space, which is not possible in a real set-up. The ratio between the signal amplitude at the sound source $x$ and microphone signal amplitudes for the left and right channels, $y_1$ and $y_2$, vary with the angle of incidence:

$$y_1(t) = g_d \cdot \cos(\alpha + 45^\circ) \cdot x(t - \tau), \quad (7)$$
$$y_2(t) = g_d \cdot \cos(\alpha - 45^\circ) \cdot x(t - \tau). \quad (8)$$

Figure 3 depicts the sensitivity magnitudes in the Blumlein XY technique. The dashed part of the plots shows where the signal is phase inverted. Both amplitude and time differences between the microphone channels determine the position of the
spatial image that a listener will perceive when both microphone signals are amplified and played through two loudspeakers in standard stereo configuration (Fig. 1). When a sound source is encircling the microphone set-up in the horizontal plane at a distance of 3 m in the frontal hemisphere ($\alpha = -90^\circ$ to $90^\circ$), the inter-channel level difference (ICLD) $\rho$ can be calculated as follows:

$$\rho(\alpha) = 20 \cdot \log_{10} \left( \frac{y_2(t)}{y_1(t)} \right)$$

(9)

$$= 20 \cdot \log_{10} \left( \frac{g_d \cdot \cos(\alpha - 45^\circ)}{g_d \cdot \cos(\alpha + 45^\circ)} \right)$$

(10)

$$= 20 \cdot \log_{10} \left( \frac{\sin(\alpha + 45^\circ)}{\cos(\alpha + 45^\circ)} \right)$$

(11)

$$= 20 \cdot \log_{10} (\tan(\alpha + 45^\circ)).$$

(12)

The results are shown in Fig. 4. Inter-channel time differences (ICTD) do not occur, because both microphone diaphragms coincide (Fig. 5). This has been frequently criticized, and it appears that the ICTDs are often confused with the interaural time differences (ITDs) that occur between the listener's eardrums, even though the underlying theory has been previously published [2].

$$y_1(t) = 0.5 \cdot g_d \cdot (1 + \cos(\alpha + 55^\circ)) \cdot x_1(t - \tau),$$

(13)

$$y_2(t) = 0.5 \cdot g_d \cdot (1 + \cos(\alpha - 55^\circ)) \cdot x_1(t - \tau).$$

(14)

The ICLD $\rho$ can be calculated for this set-up as follows:

$$\rho(\alpha) = 20 \cdot \log_{10} \left( \frac{y_2(t)}{y_1(t)} \right)$$

(15)

$$= 20 \cdot \log_{10} \left( \frac{1 + \cos(\alpha - 55^\circ)}{1 + \cos(\alpha + 55^\circ)} \right).$$

(16)

Figure 4 shows the ICLD as a function of the angle of incidence $\alpha$. Apparently, the level difference
between both microphones remains rather low compared to the XY technique. However, increasing the angle between both microphones is rather problematic, as this would result in a very high sensitivity of the set-up toward the side. Instead, both microphones are arranged with a distance between their diaphragms, for example 17 cm in the ORTF configuration (compare Fig. 2b). This way, ICTDs $\tau_\Delta$ are generated additionally. The ICTDs can be easily determined from the geometry of the set-up (compare Fig. 6):

$$\tau_\Delta(\alpha) = \frac{(r_1 - r_2)}{c_s}$$  \hspace{1cm} (17)

$$\begin{align*}
&= \frac{1}{c_s} \left( \sqrt{r_1^2 + (d/2)^2} - r_c d \cos(90 + \alpha) - \sqrt{r_2^2 + (d/2)^2} \right) \\
&= \sqrt{r_1^2 + (d/2)^2} - r_c d \cos(90 - \alpha) \right)  \hspace{1cm} (18)
\end{align*}$$

with the distance $d$ between both microphones in meters, and the speed of sound $c_s$. In Fig. 7, the ICTD is given for the ORTF set-up ($d=17$ cm) for various distances between the sound source and the center of the microphone set-up. The incoming angle of the sound wave was kept constant at $-30^\circ$. At larger distances ($r_c > 1$ m), the ICTD converges to the constant value of $-0.25$ ms. For this reason, the ICTD is often determined using the far-field approximation:

$$\tau_\Delta(\alpha) = \frac{d}{c_s} \sin(\alpha). \hspace{1cm} (19)$$

Next, the ICLDs that result from the difference between the path lengths to both microphone positions will be determined. For simplicity, it is temporarily assumed that both microphones are omni-directional rather than uni-directional. The path length for the left microphone is:

$$r_1 = \sqrt{r_c^2 + (d/2)^2} - r_c d \cos(90 + \alpha) \right)  \hspace{1cm} (20)$$

and for the right channel we find

$$r_2 = \sqrt{r_c^2 + (d/2)^2} - r_c d \cos(90 - \alpha). \hspace{1cm} (21)$$

In general, the ICLD $\rho$ is determined by applying the inverse-square law (Eq. 2):

$$\begin{align*}
\rho &= 20 \log_{10} \left( \frac{p_0 \cdot r_0}{r_1} \right) - 20 \log_{10} \left( \frac{p_0 \cdot r_0}{r_2} \right) \\
&= 20 \log_{10} \left( \frac{r_2}{r_1} \right). \hspace{1cm} (22)
\end{align*}$$

Fig. 5: Inter-channel time differences as a function of azimuth for different recording techniques.

Fig. 6: Physical relations in a two-channel near coincident microphone set-up, $M_1$ and $M_2$, to record a point source $S$. 

Fig. 7: The ICTD is given for the ORTF set-up ($d=17$ cm) for various distances between the sound source and the center of the microphone set-up. The incoming angle of the sound wave was kept constant at $-30^\circ$. At larger distances ($r_c > 1$ m), the ICTD converges to the constant value of $-0.25$ ms. For this reason, the ICTD is often determined using the far-field approximation:

$$\tau_\Delta(\alpha) = \frac{d}{c_s} \sin(\alpha). \hspace{1cm} (19)$$

Next, the ICLDs that result from the difference between the path lengths to both microphone positions will be determined. For simplicity, it is temporarily assumed that both microphones are omni-directional rather than uni-directional. The path length for the left microphone is:

$$r_1 = \sqrt{r_c^2 + (d/2)^2} - r_c d \cos(90 + \alpha) \right)  \hspace{1cm} (20)$$

and for the right channel we find

$$r_2 = \sqrt{r_c^2 + (d/2)^2} - r_c d \cos(90 - \alpha). \hspace{1cm} (21)$$

In general, the ICLD $\rho$ is determined by applying the inverse-square law (Eq. 2):

$$\begin{align*}
\rho &= 20 \log_{10} \left( \frac{p_0 \cdot r_0}{r_1} \right) - 20 \log_{10} \left( \frac{p_0 \cdot r_0}{r_2} \right) \\
&= 20 \log_{10} \left( \frac{r_2}{r_1} \right). \hspace{1cm} (22)
\end{align*}$$
The solution for Eq. 23 is shown in Fig. 8 for various distances between the sound source and the center of the microphone set-up. Again, the distance $d$ between both microphones is set to 17 cm, and the incoming angle is $-30^\circ$. At larger distances ($r_c > 1.5$ m) the ICLD converges to zero, which leads to a practical separation between ICLDs and ICTDs for coincident and near-coincident techniques. The matter becomes more complex with spaced microphone techniques, because here both the ICTDs and ICLDs are generally determined by path-length differences between the microphones and the sound sources.

In case of the ORTF set-up (and other near-coincident techniques), the distance between both microphones (17 cm) is on the same order as the distance between both eardrums, and the ICTD reaches its maximum of 1 ms, when the sound source is located sideways. The ICTD adds to the ITDs (and ILDs) at the listener’s eardrums when the recording is reproduced via two loudspeakers, and often supernatural cues, e.g. ITD magnitudes exceeding the range for natural sound sources, are observed. Nevertheless, the ICTDs extend the range of the spatial images of sources.

### 3. BINAURAL-MODEL STRUCTURE

#### 3.1. Periphery

The general structure of the binaural model is shown in Fig. 9. The transformations from the sound sources to the eardrums (influence of the outer ear and occasionally room reflections) are taken into account by filtering the sounds with HRTFs from a specific direction (e.g., $30^\circ$ azimuth for the left speaker and $-30^\circ$ azimuth for the right speaker). Afterwards, the outputs for all sound sources—typically the signals from the left and the right loudspeakers—are added together for the left and the right channel. Basilar-membrane and hair-cell behavior are simulated with a gammatone filter bank with 36 bands at a sampling frequency of 48 kHz, as described by Patterson et al. [18], and a simple half-wave rectification. To take into account that the human auditory system cannot resolve the temporal fine structure at high frequencies, the envelope of the signal is determined for frequencies above 1500 Hz by using the Hilbert transform instead of half-wave rectifying the signal.
3.2. ITD analysis

After the half-wave rectification, the normalized interaural cross correlation is estimated within each frequency band over the whole target duration $T$:

$$\Psi(\tau) = \frac{\int_{t_1}^{t_2} x_l(t) \cdot x_r(t + \tau) dt}{\left(\int_{t_1}^{t_2} x_l(t) dt \cdot \int_{t_1}^{t_2} x_r(t) dt\right)^{1/2}}, \quad (24)$$

with the internal delay $\tau$, and signals $x_l(t)$ and $x_r(t)$ in the left and right channels.

3.3. ILD analysis

An adequate method to process the ILDs in an analogous way to the processing of ITD cues is to use an array with excitation/inhibition (EI) cells. An algorithm with this characteristic was proposed by Breebaart et al. [9]. It employs an excitation-inhibition (EI) algorithm, based on the physiological findings of Reed and Blum [21]. In the investigation reported here, a version of Breebaart et al.’s algorithm was used, which was modified by the author [6] to analyze ILD cues only. In this algorithm, every cell has an excitatory and inhibitory input and is tuned to a certain ILD $\alpha$. The output of each EI cell $E(\alpha)$ is estimated as follows:

$$E_i(\alpha) = \exp\left(\frac{10^{\alpha/\text{ILD}_{\text{max}}}}{10^{-\alpha/\text{ILD}_{\text{max}}} \sqrt{P_{i,l} - \sqrt{P_{i,r}}}}\right), \quad (25)$$

with $P_{i,l}, P_{i,r}$, being the power in the left and right channels and $i$ referring to the $i^{th}$ frequency band.

In this model simulation, 81 EI-cells were used for each frequency band $i$. The ILD $\alpha$ was adjusted to values between $-40$ and 40 dB in steps of 1 dB. The EI cells are implemented in each frequency band directly after the halfwave rectification or envelope extraction.

3.4. Cross-correlation algorithm with contralateral inhibition and pre-compression

The algorithm to simulate the precedence effect for ITD analysis was adapted from the Lindemann
model [14] and previously described in [8]. The novelty of Lindemann’s algorithm was the introduction of contralateral-inhibition elements (static inhibition) into the cross-correlation model. In the model, the signals in the delay lines for the left and right channel at the corresponding delay unit. In addition to the static inhibition, Lindemann also introduced a dynamic inhibition which he defined as follows:

\[
\begin{align*}
    r(m + 1, n - 1) &= r(m, n)[1 - c_d l(m, n)] \quad (26)
    \\
    l(m + 1, n + 1) &= l(m, n)[1 - c_s r(m, n)] \quad (27)
\end{align*}
\]

with \(m\) being the index for discrete time. The variable \(n\) is the index for internal delay and \(c_s\) refers to the static inhibition constant \((0 \leq c_s < 1)\).

Now, the signals of both channels inhibit each other, thus reducing the amplitude of the signal in the opposite channel at the corresponding delay unit. In addition to the static inhibition, Lindemann also introduced a dynamic inhibition which he defined as follows:

\[
\begin{align*}
    \phi(m, n) &= c_d \cdot k(m - 1, n) + \\
    \phi(m - 1, n) \cdot e^{-T_v/T_{inh}}[1 - c_d \cdot k(m - 1, n)].
\end{align*}
\]

with \(\phi\) being the running, dynamic inhibition function. The variable \(c_d\) is the dynamic inhibition constant \((0 \leq c_d < 1)\). \(T_v\) is the time delay of a delay unit \((21 \mu s = 1/f_s)\), and \(T_{inh}\) represents the fade-off time constant of the nonlinear low pass.

Originally, Lindemann had only analyzed narrow-band signals, and thus, no bandpass filter bank was required in his investigation. In our analysis, the gammatone-filter bank and halfwave rectification that were described in Section 3.1 are used in the model.

One characteristic of the Lindemann model is that the effective degree of inhibition depends on the signal’s amplitude. To avoid the degree of inhibition being much lower in those frequency bands with less signal energy, the signal’s maximum in each band was scaled to one. After the half-wave rectification, the cross-correlation patterns were computed and multiplied with the average power of the stimulus in the left and right channels measured in that frequency band.

Another feature of the Lindemann model is the combined analysis of ITDs and ILDs. A side effect of the contralateral inhibition is a shift of the cross-correlation peak toward the channel with the higher energy. Since it is one of the goals to investigate the influence of ITD and ILD cues separately, it is better to analyze both cues in separate algorithms. It was therefore decided to modify the Lindemann algorithm in such a way that it is almost independent of ILDs. Fortunately, the algorithm’s dependence on ILDs is quite low, and in fact Lindemann had to introduce monaural processors to enhance the influence of ILD cues. Besides omitting the monaural processors, the signal was compressed after the halfwave rectification by taking the signal to the power of 0.25 before it was scaled. In this way, the influence of the ILDs is reduced furthermore, as can be seen in Fig. 10. The settings of the model were therefore decided to modify the Lindemann algorithm for the processing of ILD cues, the EI model, as was described in Section 3.3 was installed as a running algorithm with inhibition units. Therefore, Equation 25 had to be modified to:

Fig. 10: Demonstration of the compression algorithm that was introduced in the Lindemann model to reduce the influence of the ILDs for a noise burst (500-Hz center frequency, 100-Hz bandwidth, 0-µs ITD). The solid line shows the average cross-correlation function, when the ILD of the signal is zero, and the dotted line, when the ILD of the signal is 20 dB (no compression). After inserting the compression stage, the peak moves—despite an ILD of 20 dB—almost to the center (dashed line).
\[ E_i(m, \alpha) = \exp \left( 10^{\alpha/10L_{\text{max}}} \sqrt{P(m)_{i,t}} - 10^{\alpha/10L_{\text{max}}} \sqrt{P(m)_{i,r}} \right) \]  

(29)

with \( P(m)_{i,t}, P(m)_{i,r} \), the power in the left and right channels, and \( i \) and \( m \) referring to the \( i \)th frequency band and the \( m \)th time slot. Before the outputs of the halfwave rectification were sent to the inputs of the EI algorithm, they were convolved with a Hanning window of 10-ms duration to acknowledge the effect of binaural sluggishness. Afterwards, the outputs of the EI algorithm were down-sampled to a resolution of 1 ms.

The inhibition function \( E_{\text{inh}}(m) \) and the new, inhibited function \( E_{\text{new}}(m) \) were calculated iteratively as:

\[ E_{\text{inh}}(m, \alpha) = \frac{\max(E_{\text{new}}(m - 1, \alpha))}{E_{\text{new}}(m - 1, \alpha)} \cdot c_1 + E_{\text{inh}}(m - 1, \alpha) \cdot c_2 \]

(30)

\[ E_{\text{new}}(m, \alpha) = \frac{\max(E(m, \alpha)) - E_{\text{inh}}(m, \alpha)}{E(m, \alpha)} \]

(31)

with \( c_1 \) and \( c_2 \) being two inhibition constants (0 \( \leq \) \( c_1, c_2 < 1 \)). After having calculated each step, negative values of \( E(m)_{\text{inh}} \) and \( E(m)_{\text{new}} \) were set to zero, as a negative activity of the cells would be invalid for a physiologically oriented model. The settings of the model were kept the same as described in [8].

3.6. Remapping

For broadband signals, it is useful to remap the cross-correlation functions from interaural time differences to azimuth positions. Otherwise, the peaks of the cross-correlation functions will not necessarily line up at one lag for a single sound source because the ITDs of the HRTFs are frequency dependent. To calculate the ITDs of the HRTFs throughout the horizontal plane, the HRTF catalog was measured in a resolution of \( 5^\circ \) in the horizontal plane. The measurement procedure is described in [4]. After filtering the HRTFs with the gammatone filter bank, the ITDs for each frequency band and angle are estimated using an interaural cross-correlation (ICC) algorithm. This frequency-dependent relationship between ITDs and azimuth angles is used to remap the output of the cross-correlation stage (ICC curves) from a basis of ITDs \( \tau(\alpha, f_i) \) to a basis of azimuth angles in every frequency band \( i \):

\[ \tau(\alpha, f_i) = g(\text{HRTF}_{\text{l}}, \text{HRTF}_{\text{r}}, f_i) \]

(32)

\[ \tau(\alpha, f_i) = g(\alpha, f_i) \]

(33)

with \( \alpha = \) azimuth, \( \theta = \) elevation = \( 0^\circ \), \( r = \) distance = 2 m, \( \text{HRTF}_{\text{l/r}} = \text{HRTF}_{\text{l/r}}(\alpha, \theta, r), f_i = \) center frequency of bandpass filter.

Next, the ICC curves \( \psi(\tau, f_i) \) are remapped to a basis of azimuth angles using a simple \texttt{for}-loop in Matlab:

```
for alpha=0:5:360
    psi_rm(alpha,freq)=psi(g(alpha,freq),freq);
end
```

An analogous method is used to remap the output of the Excitation/inhibition cell array from ILDs to azimuth angles.

In Fig. 11, one-dimensional examples of such spatial maps are depicted. In the left panel, the relationship between the ITDs and the azimuth in the horizontal plane is shown for three different frequency bands. In the right panel, the relationship between the ILDs and the azimuth is illustrated. Given the usage of such maps, the output of the cross-correlation algorithm on the basis of ITDs can be re-mapped on the basis of the azimuth in the horizontal plane as shown in Fig. 12. Ambiguities often occur as mentioned in the introduction section. For example, as seen in Fig. 12, the ITD-based analysis cannot reveal if the sound was presented from the front or the rear hemisphere. In former approaches ([16], [20]), the model will have to decide for one of the two equally high peaks.

3.7. Decision device

In the decision device, the position and the apparent width of the auditory event has to be determined. In principle, three cues are known that all have influence on the apparent source width:

1. interaural coherence in each frequency band
2. the lateral mismatch of the azimuth-mapped peak positions across all frequency bands
3. the variation of the lateral peak positions in each frequency band over time

In the current implementation, the decision device of the binaural model described here, makes use of the interaural coherence and the lateral mismatch of the peak positions. Previous model algorithms exist that predict the width of the auditory image based on the interaural coherence [2] or interaural time difference fluctuations [10], [22], [23], [24]. In this work, the interaural coherence is considered but not the fluctuations of binaural cues. For the ITDs, the remapped normalized cross correlation function for each frequency band \( i \) is multiplied with the estimated sound pressure level \( \gamma_i \) in this band (the sound pressure level for the left and right channels are added for this purpose). If the interaural signals are fully correlated (coherence of one), the peak height is equal to \( \gamma_i \), if the signal is partly decorrelated the signal becomes smaller than \( \gamma_i \). Negative values are not observed since the signals have been half-wave rectified, or the signals' envelopes have been extracted before cross-correlating them. Afterwards, all cross-correlation functions are summed up for the frequency bands 1–16 (fine-structure analysis), and divided by the sum of all sound pressure levels:

\[
\Psi_f(\alpha) = \frac{\sum_{i=1}^{16} \Psi_i(\alpha) \cdot \gamma_i}{\sum_{i=1}^{16} \gamma_i}.
\]

(34)

The same is done for the cross-correlation functions that were determined from the envelope signals (bands 17-36):

\[
\Psi_e(\alpha) = \frac{\sum_{i=17}^{36} \Psi_i(\alpha) \cdot \gamma_i}{\sum_{i=17}^{36} \gamma_i}.
\]

(35)

A similar function is determined for the ILD analysis:

\[
E(\alpha) = \frac{\sum_{i=1}^{36} E_i(\alpha)}{36}.
\]

(36)
Of course, the calculation of the coherence does not apply to level differences. The position and maximum peak height indicates the position and the apparent source width for each of the three analyzed cues: interaural time differences (fine structure and envelopes) and interaural level differences. Multiple peaks will appear if the output of the model is ambiguous. All three functions can be integrated for an overall estimate $F$:

$$F(\alpha) = \frac{1}{3} \left( \frac{1}{16} \psi_e(\alpha) + \frac{1}{20} \psi_e(\alpha) + \frac{1}{36} E(\alpha) \right).$$

In this approach, the information in each frequency band is weighted equally. In future, psychoacoustic weighting functions have to be derived from listening experiments. Such weighting function have been already derived for other psychoacoustic tasks [11], [26], [25]. In this study, however, the evaluation of the general model structure is more important than simulating the auditory system in every detail. For complex psychoacoustic tasks such as localization and determining the apparent source width, large variations in weighting individual cues are expected across listeners.

4. STIMULI

Before reporting on the evaluation of the model algorithms that were introduced in the previous section, the test material to evaluate the model will be described briefly. The noise bursts (200-ms duration, 20-ms cos²-ramps) serving as a sound source were generated digitally at a sampling frequency of 48 kHz and 16-bit resolution. In the initial part of this investigation, the transfer functions between the sound source and each microphone was calculated according to the theory described in Section 2. To simulate the whole pathway between sound source and the signals at the eardrums, the sound source was filtered with the transfer function for each microphone. Afterwards, each of the two microphone signals was filtered with the HRTFs for the corresponding loudspeaker position (left microphone: 30°, right microphone: −30°). For this purpose, the HRTFs were taken from the same catalog that was used to remap the cross-correlation and EI-cell functions. These signals were summed up for the left and then for the right ear and analyzed using the binaural model. A reference condition with a direct pathway from the sound source to both ears was simulated as well, by filtering the source signal with the HRTFs of the corresponding position.

At a later stage of this research, impulse responses of real microphone set-ups were measured to replace the theoretical ones. Apart from this, the procedure described above was kept. The measurements were conducted in Clara Liechtenstein Hall, a recital hall at McGill University with moderate reverberation times. The following microphones were used in the measurements (Sennheiser MKH 30, Blumlein technique; DPA 4011, ORTF technique; AKG C-414, MS technique; DPA 4007, spaced-omni technique). The measurement software was executed from a personal computer (Pentium 4 with Motu 896 sound card). A loudspeaker (Genelec 1030a) was used as a sound source.

5. MODEL-SIMULATION RESULTS

5.1. Results for a target at 30° azimuth

Figure 13a shows the model results for a target position at 30°. In this setting, the sound source was analyzed directly by the model in absence of a microphone set-up (reference condition). The upper panel shows the ICC for every frequency band for different internal delay times $\tau$. The height of the peak in every frequency band shows the coherence in this band.

The color-coding of the values is shown in the color map besides the color plot. The ICC peaks become much broader for the frequency bands 17 (1500 Hz) and above, because the envelope and not the fine structure of the signal was analyzed here. In the lower panel, the ICC function averaged over the frequency bands (1–16) is depicted in blue. In this case the peak is fairly narrow and located at 30° azimuth corresponding to the position of the source. Also for the analysis of the envelopes in the higher frequency bands (17–36) the peak is located at the same position, but the width of the peak is much larger due to the circumstances described above.

Figure 13b shows the analysis for the Blumlein microphone configuration. Again, the source was positioned at 30° azimuth, but the cross-correlation peaks appear at lower internal delays than in the
reference condition. Hence, the model analysis for ITDs predicts that the spatial image would be perceived more toward the center line than it was the case for our original sound source at 30°. The graphs for the remaining two coincident microphone techniques: coincident cardioid (Fig. 13c) and MS (Fig. 13d) show a very similar pattern to the Blumlein technique. A different result is obtained if the microphones are separated in space. While those differences are relatively small for the ORTF technique, the cross-correlation peaks in the spaced-omni case are found to be at much higher internal delays \( \tau \), especially for center frequencies above 500 Hz. Below this critical frequency, the cross-correlation peaks are located closer to the centerline, because the contralateral signals are hardly attenuated by the listeners head (no ILDs), and the cross-talk from a loudspeaker to the ear at the opposite side decreases the measured ITD.

Figure 14 shows the results for the ILD analysis for the same conditions as in the ITD analysis. The reference condition depicts the ILDs that occur for a target source at 30° (Fig. 14a). Below a center frequency of 500 Hz, the ILDs have values close to zero as indicated by the red-colored peaks of the EI-cell functions. Above this value moderate ILDs between 5 and 10 dB can be observed. Only at frequencies around 9000 Hz, this value increases to approximately 20 dB.

For the three coincident techniques the location of the EI cells peaks are fairly close to the reference condition (Fig. 14b–d). In general, the measured ILDs are slightly smaller than for the reference condition with some exceptions, for example, the high ILD value of approximately 20 dB that is found for the coincident-cardioid set-up at a center frequency of 2 kHz. The ILDs are smaller for the ORTF set-up (Fig. 14e) and for low frequencies even negative ILD values are measured. The small recording angle (110° vs. 130° in the coincident-cardioids condition) explains why the ILDs are lower in general, but it does not explain the negative values. The reason for the latter are interference effects that are induced by the inter-channel time differences that characterize non-coincident techniques (A similar effect has been described in [7]). The interference effect become more prominent in the spaced-omni set-up (Fig. 14f). For frequencies below 1.5 kHz unnaturally large ILD values are observed in both directions.

In the next analysis step, the data that were previously shown in Fig. 13 are remapped to azimuth angles before they are visualized. In the lower panel of each graph, the ICC function averaged over the frequency bands (1–16) is depicted in blue. For the reference condition, the peak is fairly narrow and located at 30° corresponding to the position of the source. We also find a peak at the corresponding rear position (150°). In general, a second peak is always observed in the ITD analysis, but in the following lines this work concentrates on the analysis within the frontal hemisphere. Also for the analysis of the envelopes in the higher frequency bands (17–36) the peak is located at the same position, but the width of the peak is much larger due to the analysis of the envelopes instead of the carrier signal as was described in Section 3. For the three coincident techniques, the peaks are located at approximately 20° instead of 30°, but the peaks line up well at this angle (Fig. 15b–d). For the ORTF set-up, on the other hand, the average peak position is closer to the peak position in the 30° reference condition (Fig. 15e). However, the peaks do not line-up at one position anymore. Instead, the position of the peaks shifts more outwards with increasing frequency. This feature becomes more apparent for the spaced-omni technique. Here, the peaks above 500 Hz are located at 90° (Fig. 15f).

The remapped ILD data is shown in Fig. 16. For the reference condition, all maxima line up at 30° as expected (Fig. 16a). In this case, the peak for the 30° condition is higher than for the corresponding rear position, because ILDs provide better cues for resolving front/back directions than ITDs. More variation is found for the three coincident techniques, and similar to the ITD analysis the average peak location is found at 20° rather than at 30° (Fig. 16b–d). For the coincident-cardioids technique the average peak position is even below 15°. After the stereo reproduction of our microphone signals, the peaks for the front and corresponding rear position are aligned in height. Thus, the information that the signal was presented in front of the listener is not coded, despite the fact that both loudspeakers are located in front of the listener. Typically, front/back confusions do not occur in stereophonic
displays and the auditory image is usually in front of the listener. Head movements, that are not captured in this model simulation yet, would easily allow to resolve front/back confusions, and of course real listeners are typically aware of the loudspeaker locations.

For the two non-coincident techniques, the averaged peaks are located between 0° and 10° (Fig. 16e+f), but the variation is much larger for the spaced-omni technique than for the ORTF set-up.

5.2. Results across different target positions
After analyzing the case for a target located at 30° azimuth, the model analysis for other directions is described in this section. For sound recording purposes, the analysis of sound sources in the frontal horizontal plane would have been sufficient to cover many aspects. Nevertheless, for each microphone technique the azimuth angles between −180° and 180° in steps of 5° will be displayed to show the model’s strength in processing multiple peaks. For the given directions, the model should be able to predict the lateral placement of the auditory events, which will lead to the prediction of the localization curve. In addition, the model should also give some indications regarding the perceived lateral extent of the auditory events.

Figure 17 shows the data for the fine-structure ITD analysis (frequency bands 1–16). Each graph represents the results for a different microphone technique. The top left graph shows the results for the reference condition (Fig. 17). Along the x-axis the azimuth angles of the target source is displayed. The y-axis shows the activity of the interaural cross-correlation function mapped to azimuth angles. In these plots, the activity patterns were integrated over the patterns in each frequency band as was shown in Fig. 15 for the 30° case. Basically, the lower panels in Fig. 15 show the cross section of this plot at 30° (x-axis). For the reference condition, the maximum peaks are located in the ascending diagonal, which means that the model predicts the position of the target accurately. At both lateral ends (−90° and 90°) the curve widens slightly, because the there is not much ITD variation across the outer angles.

The model analysis also depicts information about the spatial extent of the sound source as described in Section 3.7. For the reference condition, the peak maximum reaches almost the value of one for all directions, which means that the peaks are well aligned across frequency bands indicating a small apparent source width. Also for the Blumlein set-up (Fig. 17b), the peak maxima have a value of one for most directions. In a typical recording situation the primary sources (e.g., musical instruments) would be positioned within an angle of −45° to +45°. Within this range the localization curve is fairly straight, but less steep than the diagonal shown in Fig. 17a, indicating that the directions are compressed when reproduced in a standard stereo loudspeaker set-up (This depression would vanish if the loudspeakers are shifted from ±30° to ±45°, but then there is the risk of having a “hole in the middle”). At ±45°, the target is on axis for one microphone and off-axis for the other microphone. Nevertheless, the peak shifts further outward for larger angles. This phenomenon can be easily explained by the inverse phase of the rear lobe of a bidirectional microphone (If the rear lobe would be in phase, the position of the maximum peak would return to 0° with increasing lateral angle). It is very absorbing that by using the Blumlein technique, ITD cues outside the range of the phantom image field spanned by both loudspeakers can be generated. At the outer angles (> |75°|) the peak height decreases and the curve becomes wider. In these conditions, both lobes have roughly the same sensitivity but are out of phase. The ITDs are therefore determined by the center frequency of the frequency bands (which do not line up at one lag) rather than the position of the sound source. The remaining two coincident techniques (MS and coincident cardioid, Fig. 17c–d) show a decompressed localization curve as well, but the straight curve continues toward the outer angles, because both techniques “effectively” do not have a rear lobe which is out of phase. Due to the space between both microphone capsules, the localization curve is steeper for the ORTF technique within a range of −40° and +40°. In this range, the curve is even steeper than the diagonal of the reference condition, indicating a decompressed localization curve (it will be shown later that this decompression effect is partly compensated by the ILD cues). At the outer angles (> |40°|), the localization curve is flat. For these angles no further shift of the auditory events is expected from analyzing the ITD cues. The
localization curve for the spaced-omni technique is by far steeper than it was for the remaining microphone techniques. Already at sound source angles of ±30°, the ITD cues coded in the spaced-omni technique correspond to source angles of ±90°. For angles larger than ±30°, the ITD cues exceed values that are generally measured in HRTFs. However, side peaks occur within the analyzed range of ±1 ms, but these do not line up at one lag. This explains the low activity (≈ 0.6) for angles above ±45°.

The analysis of the envelopes’ ITDs (frequency bands 17–36, Fig. 18) confirms most findings that were made for the ITD analysis of the fine structure signal (frequency bands 1–16, Fig. 17). Due to the high correlation across the range of the internal delays τ, the lower border of the color plot was set to 0.8 instead of 0.0. A noteworthy case is the pattern for the Blumlein technique. When analyzing the ITDs of the envelopes, the out-of-phase characteristic of the rear lobe of a bidirectional microphone has no effect, and in this case the localization curve moves toward the center line for angles above ±45°.

Major differences between the tested microphone techniques exist also for the coding of ILD cues (Fig. 19). All three coincident techniques show an accurate reproduction of the localization curve for ILD cues within the range of the recording angle (Fig. 19b–d). The localization is slightly compressed though, with a similar compression factor to the corresponding ITD-based localization curves. Outside the recording angles, the position of the activity pattern remains either constant (coincident cardioid set-up) or moves back toward the center line with increasing angle (Blumlein technique, MS technique). The localization curve obtained with the ORTF technique is relatively flat, not very straight, but increases with increasing sound source angle throughout the whole lateral field (−90° to 90°).

The ILDs cues of the average activity pattern never exceed a range of −30° to 30°. Large ILD cues were previously found for frequency-band wise analysis of the spaced-omni technique (Figs. 14f and Figs. 16f).

On average, however, these ILD cues only activate the region around 0°, independent of the angle of incidence.

The combination of all three cues (ITD fine structure cues, ITD envelope cues, and ILD cues) leads to the final model prediction of the auditory event’s lateral position (Figs. 20). In case of the reference condition, all cues line up perfectly well and the main activity zones correlate with the target angles (Figs. 20a). Within the recording angle (≈ ±45°), all cues lead into the same direction for the three coincident microphone techniques (Figs. 20b–d) and the peak heights of the overall activity patterns are found to be in between the peak heights for the ILD and ITD based analyses (Figs. 21b–d). Noteworthy are the activity patterns for the Blumlein technique outside the recording angle (> | ±45°|). Here the ILD cues (Fig. 19b) and the ITD fine structure cues (Fig. 17b) lead into opposite directions which decreases the peak heights of the combined activity patterns. In general the average peak height is lower for the ORTF technique then for the coincident techniques, because ITD and ILD cues do not correspond as well to each other as for the coincident microphone techniques (Figs. 20e and 21e). The lowest overall peak height was found for the spaced-omni technique (Figs. 20f and 21f). Here, ILD and ITD cues do not line up at all. While the localization curve for the ITD cues was extremely steep within the recording angle (Fig. 17f), the localization curve for the ILD remained to be close to zero for all tested angles of incidence (Fig. 19f).

5.3. Investigating realistic scenarios

Figure 22 shows the results of the cross-correlation model if measured impulse responses are used to investigate microphone techniques instead of simulated ones (left panels). Now, the cross correlation peaks line up less well than it was the case for the theoretical microphone set-ups. For the 0° conditions (spaced omni, top panels; Blumlein, bottom panels), the heights of the averaged peaks drop from nearly 1.0 in the theoretical condition (Fig. 21), to 0.7. This phenomenon is due to both the misalignment of the peaks and the reduced coherence in the reverberant environment. The performance of the model improves, when an inhibition mechanism is used to simulate the precedence effect (right panels). In this simulation, however, the integration of coherence is not possible.

Figure 23 shows the result of the ILD analysis. Again, the presence of reverberation, misaligns the peaks (left panel). Unfortunately, the inhibition stages do not resolve this problem (right panels). At
this point, it remains unclear whether the structure of the EI-cell array has to be improved or whether we observe the same effect in the auditory system. It would also be worth to investigate whether the differences between both model performances are greater, if a concert space with larger reverberation times than Clara Liechtenstein Hall is selected.

6. DISCUSSION AND OUTLOOK
In general, the proposed model structure enables the analysis of microphone techniques. The model structure has less difficulties in handling multiple peak phenomena than previous model approaches. In future it is planned to better tune the model to psychoacoustic data. It will be necessary to find frequency weighting curves to balance the importance of the individual cues. It is also planned to analyze microphone set-ups in different concert spaces to gain more insight into why different microphones set-ups are preferred in different halls.

7. ACKNOWLEDGEMENT
This investigation was supported by Grants of the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Government of Québec (VRQ). I would like to thank Wieslaw Woszczyk and William L. Martens for their support and the helpful discussions.

8. REFERENCES


Fig. 13: Cross-correlation activity patterns for a target at 30° using different microphone techniques: (a) reference condition, (b) Blumlein, (c) MS, (d) coincident cardioids, (e) ORTF, (f) spaced-omni.
Fig. 14: EI-cell activity patterns for a target at $30^\circ$ using different microphone techniques: (a) reference condition, (b) Blumlein, (c) MS, (d) coincident cardioids, (e) ORTF, (f) spaced-omni.
Fig. 15: Same as Fig. 13, but for cross-correlation functions that were remapped to azimuth angles.
Fig. 16: Same as Fig. 14, but for EI-cell functions that were remapped to azimuth angles.
Fig. 17: Average results of the cross-correlation model for different target positions (fine-structure analysis, frequency bands 1–16). The following microphone techniques were simulated: (a) reference condition, (b) Blumlein, (c) MS, (d) coincident cardioids, (e) ORTF, (f) spaced-omni.
Fig. 18: Average results of the cross-correlation model for different target positions (signal-envelope ITD analysis, frequency bands 17–36). The following microphone techniques were simulated: (a) reference condition, (b) Blumlein, (c) MS, (d) coincident cardioids, (e) ORTF, (f) spaced-omni.
Fig. 19: Average results of the EI-cell model for different target positions (ILD analysis, frequency bands 1–36). The following microphone techniques were simulated: (a) reference condition, (b) Blumlein, (c) MS, (d) coincident cardioids, (e) ORTF, (f) spaced-omni.
Fig. 20: Average results of the binaural model for different target positions (combined ITD and ILD analysis, frequency bands 1–36). The following microphone techniques were simulated: (a) reference condition, (b) Blumlein, (c) MS, (d) coincident cardioids, (e) ORTF, (f) spaced-omni.
Fig. 21: Peak heights for the average results of the binaural model for different target positions for the data shown in Figs. 17–20. The following microphone techniques were simulated: (a) reference condition, (b) Blumlein, (c) MS, (d) coincident cardioids, (e) ORTF, (f) spaced-omni.
Fig. 22: Cross-correlation activity patterns for different conditions: (a) spaced-omni, 0° target, without inhibition, (b) spaced-omni, 0° target, with inhibition, (c) spaced-omni, 90° target, without inhibition, (d) spaced-omni, 90° target, with inhibition, (e) Blumlein, 0° target, without inhibition, (f) Blumlein, 0° target, with inhibition.
Fig. 23: Same as Fig. 22, but for the ILD analysis using EI cells.