A Comparison of Computational Precedence Models for Source Separation in Reverberant Environments*

CHRISTOPHER HUMMERSONE,[1] *AES Member,* **RUSSELL MASON,**[1] *AES Member,* **AND**
(c.hummersone@surrey.ac.uk) (r.mason@surrey.ac.uk)
TIM BROOKES,[1] *AES Member*
(t.brookes@surrey.ac.uk)

[1] *University of Surrey, Guildford, UK*

Reverberation continues to be problematic in many areas of audio and speech processing, including source separation. The precedence effect is an important psychoacoustic tool utilized by humans to assist in localization by suppressing reflections arising from room boundaries. Numerous computational precedence models have been developed over the years and all suggest quite different strategies for handling reverberation. However, relatively little work has been done on incorporating precedence into source separation. This paper details a study comparing several computational precedence models and their impact on the performance of a baseline separation algorithm. The models are tested in a range of reverberant rooms and with a range of other mixture parameters. Large differences in the performance of the models are observed. The results show that a model based on interaural coherence and onset-based inhibition produce the greatest performance gain over the baseline algorithm. The results also show that it may be necessary to adapt the precedence model to the acoustic conditions of the room in order to optimize the performance of the separation algorithm.

0 INTRODUCTION

Computational separation of mixtures of sound is an area of high research interest due to the numerous applications for separation algorithms, including front-end processing for missing data speech recognition and enhancement of hearing prostheses and communication devices such as mobile phones. Many algorithms have been proposed that utilize a variety of processing techniques that work well in anechoic conditions. However, in many situations reverberation is likely to be present, and unfortunately it continues to be a major obstacle for separation algorithms due to its corruption of many of the acoustical cues on which these algorithms rely. Similarly, a number of localization algorithms have been proposed that work well in anechoic conditions (e.g., [1]), but these often perform poorly in reverberant environments because interaural cues are also corrupted by reverberation. Consequently, reducing the detrimental effects of reverberation continues to be an important research goal not only for researchers in source separation but also for researchers working in other areas of signal processing.

A number of models have been suggested that attempt to reduce the deleterious effects of reverberation using engineering-based methods. For example, the Wiener filter is shown in [2] to be effective at reducing the effect of reverberation on sound source separation. In contrast, numerous human psychophysical and perceptual mechanisms for suppressing the effects of reverberation are well documented, which have occasionally provided a source of inspiration for researchers in signal processing. One such mechanism is the precedence effect.

The precedence effect (for a review see [3]) is described in the perceptual literature as being an important mechanism for enhancing our ability to localize sounds in reverberant environments. Often referred to as the "law of the first wave front," the precedence effect describes an auditory mechanism that is able to weight the first (direct) wavefronts of a sound over later wavefronts arriving as reflections from other surfaces. However, relatively little work has been carried out on incorporating precedence effect processing into separation algorithms that utilize spatial cues. Work carried out so far is based on that of Palomäki et al. [4] (see also [5]). However, as Palomäki et al. note, the precedence model they utilize is somewhat simplified and further work could be done in order to improve its localization capabilities.

[*] This paper is a revised and expanded version of AES 128th convention paper 7981

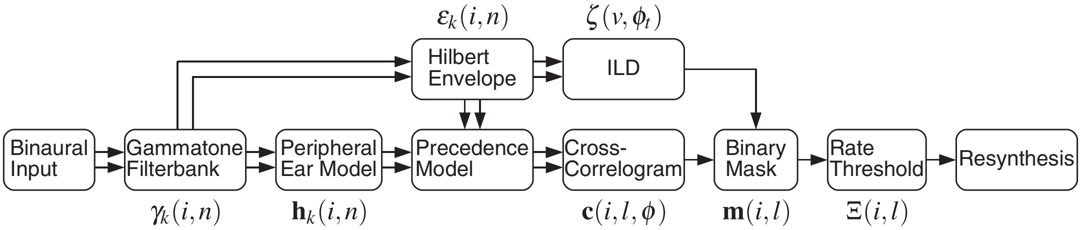

Fig. 1. Schematic of the baseline separation algorithm and precedence model based on [4].

The aim of this paper is to investigate whether an enhanced precedence model can improve the separation performance of a baseline separation algorithm. Numerous computational precedence models have been proposed in the literature (see for example [6,7,8,9,10]). A comprehensive study of these precedence-based processing strategies was conducted and their impact on the performance of the baseline separation algorithm was investigated. This baseline algorithm is described in Section 1, the additional models are described in Section 2, the experimental procedure is given in Section 3, results are presented and discussed in Section 4, and the findings are concluded in Section 5. The human auditory system uses many cues in order to separate the constituent signals of a mixture of sounds [11]. To facilitate investigation of the precedence effect, however, this paper will consider spatial cues only.

The separation algorithm and precedence models described in this paper have been made available as Matlab code at the following URL: http://iosr.surrey.ac.uk/software. The binaural room impulse responses used in the experimental section are also available to download from this URL.

1 THE BASELINE ALGORITHM

This section will first describe the baseline separation algorithm (Section 1.1), which is heavily based upon the aforementioned work of Palomäki et al. [4] (note: although every attempt has been made to follow the principles of this algorithm, due to practicalities of implementation and modifications required to enable the evaluation method described below, the processing utilized is not identical). The work includes a simple precedence model, described in Section 1.1.2. The architecture of the baseline algorithm is summarized in Fig. 1.

1.1 The Baseline Separation Algorithm

As shown in Fig. 1, the baseline algorithm takes a binaural input and begins its processing with a gammatone filterbank and peripheral ear model.

In reverberant environments, the correlation between the left and right ear signals is likely to be lower than that under anechoic conditions, potentially having a negative impact on the algorithm's localization and separation performance. To minimize this impact, a precedence model is incorporated that inhibits or suppresses information that is likely to be corrupted by reverberation. The model aims to achieve this by retaining onsets and suppressing information that follows them. The precedence model operates on the output of the peripheral ear model and calculates the cross-correlogram for each frame and frequency channel.

The cross-correlograms are then warped to the azimuthal domain and used to estimate the relative strengths of two competing signals arising from spatially-separate sound sources. These relative strengths are used to calculate a binary mask, each element of which is set to one when the correlation at the target source azimuth is greater than the correlation at the interfering source azimuth. The binary mask is used to perform the separation, effectively acting as an array of amplitude envelopes applied to the outputs of the filterbank.

Each component shown in Fig. 1 is described in the following sections.

1.1.1 Gammatone Filterbank, Hilbert Envelope, and Peripheral Ear Model

As shown in Fig. 1, the binaural left and right signals are first passed through a fourth-order gammatone filterbank [12] to simulate cochlear frequency selectivity (32 channels are employed, in the range 50–7500 Hz, equally spaced on the ERB-rate scale). The outputs of the gammatone filterbank are then half-wave rectified as a crude model of the Inner Hair Cells (IHCs); the results are denoted $\mathbf{h}_{\mathrm{L}}$ and $\mathbf{h}_{\mathrm{R}}$. The Hilbert envelopes ε_k (for ear $k \in \{\mathrm{L}, \mathrm{R}\}$) of each of these signals are used to estimate the auditory nerve firing rate $\mathbf{u}_k$ at time frame l:

$$\mathbf{u}_k(i,l) = \acute{\varepsilon}_k\big(i,(l-1)M+1\big)^{0.3} \tag{1}$$

where

$$\acute{\varepsilon}_k(i,n) = \varepsilon_k(i,n) - e^{-\alpha_s}\acute{\varepsilon}_k(i,n-1), \tag{2}$$

M is the frame length in samples (10 ms), $\mathbf{u}$ denotes the auditory nerve firing rate, and α_s is a time constant set in samples to 8 ms. The precedence model is then introduced to inhibit the inner-hair-cell-modeled data.

1.1.2 The Baseline Precedence Model

Many computational precedence models (especially those implemented in this paper) suggest that precedence is achieved by an inhibitory mechanism in the auditory periphery, rather than by a cognitive mechanism. The baseline precedence model conforms to this modus operandi by attempting to inhibit inner-hair-cell-modeled data after each onset that are likely to be corrupted by reverberation. The inhibited data are used to calculate the cross-correlograms.

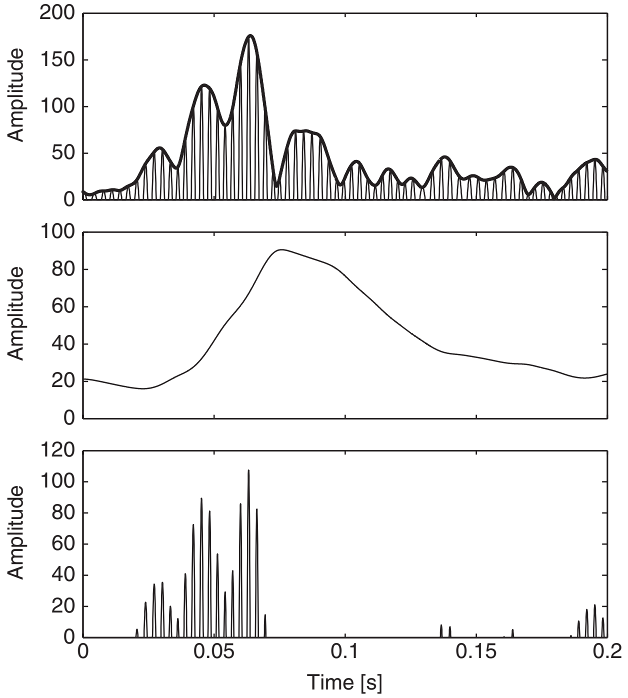

Fig. 2. Examples of the processing in the baseline precedence model (this excerpt is used in subsequent examples for other models). *Top:* Half-wave rectified gammatone filter output (302 Hz frequency channel) showing the fine structure and Hilbert envelope. *Middle:* The onset-de-emphasised low-pass filtered signal envelope. *Bottom:* The inhibited fine structure.

In the implementation, the baseline model employs an onset-de-emphasising low-pass filter with an impulse response of the form:

$$h_{\text{lp}}(n) = Ane^{-n/\alpha_p} \tag{3}$$

where α_p is a time constant chosen to be the number of samples corresponding to 15 ms and A is set to give unity gain at DC. This is used to filter the Hilbert envelope ε_k to produce an "inhibitory signal." This inhibitory signal is then subtracted from the half-wave rectified gammatone filterbank fine structure. The process is summarized in the following way:

$$\mathbf{r}_k(i,n) = \max\Big(\mathbf{h}_k(i,n) - G\big(h_{\text{lp}}(n) * \varepsilon_k(i,n)\big), 0\Big) \tag{4}$$

where G is an inhibitory gain factor that is set to 1. The precedence-modeled fine structure $\mathbf{r}$ is used to obtain the cross-correlograms (see (5)). An example of the inhibition procedure is shown in Fig. 2. As can be seen in the figure, the onset-de-emphasised low-pass-filtered envelope (c) has been subtracted from the fine structure (b), thus retaining the onset (d). The information following the onset, that is likely to be corrupted by reverberation, has been suppressed.

Zurek [13] notes that inhibited information is only used in localization and that reverberation makes a significant contribution to the timbral and spatial characteristics of a perceived sound. The baseline algorithm reflects this by only using precedence-modeled information in the localization aspect of the algorithm.

1.1.3 Cross-Correlograms

The cross-correlograms $\mathbf{c}$ for each frame are obtained by cross-correlating the precedence-modeled fine structure $\mathbf{r}_k$ over a three-frame rectangular window:

$$\mathbf{c}(i,l,\tau) = \sum_{d=0}^{3L-\tau-1} \mathbf{r}_{\text{L}}\big(i,(l-1)L+d+\tau\big)\mathbf{r}_{\text{R}}\big(i,(l-1)L+d\big) \tag{5}$$

where τ denotes the discrete lag (representing Interaural Time Difference (ITD)) of the cross-correlation such that $\{\tau \in \mathbb{Z} : -T \leq \tau \leq T\}$, $T = 1$ ms (in samples) and $\mathbb{Z}$ is the set of integers.

The data from the cross-correlograms are subsequently warped from ITD to azimuth to yield $\mathbf{c}(i,l,\phi)$, where ϕ denotes azimuthal angle such that $\{\phi \in \mathbb{Z} : -90° \leq \phi \leq 90°\}$, since the relationship between ITD and azimuth is frequency-dependent [14]. The warping function is derived from Kuhn's [14] work. Specifically,

$$\text{ITD} = \frac{\Pi\eta\sin\phi}{c_0} \tag{6}$$

where Π varies with frequency f (in Hz) such that

$$\Pi = \begin{cases} 3 & f \leq 500 \\ 2.5 + 0.5\cos\left(\pi\dfrac{\log_2\frac{\sqrt{6}f}{1250}}{\log_2 6}\right) & 500 < f < 3000 \\ 2 & f \geq 3000 \end{cases} \tag{7}$$

where c_0 is the speed of sound (344 ms^{-1}) and η is the effective radius of the head, which Kuhn derives as 0.093 m, somewhat larger than typical skull perimeter measurements, perhaps due to protruding features such as the nose and pinnae. Since Kuhn is not specific about the change in Π between 500 and 3000 Hz, a raised cosine function is chosen to vary Π "smoothly."

The azimuthal-domain cross-correlograms are then transformed to skeleton cross-correlograms [4,15] in the following way:

$$\mathbf{s}(i,l,\phi) = \mathbf{q}(i,l,\phi) * \exp\left(\frac{-\phi^2}{2\sigma^2(i)}\right) \tag{8}$$

where

$$\mathbf{q}(i,l,\phi) = \begin{cases} \mathbf{c}(i,l,\phi) & \begin{array}{l}\text{if}\big(\mathbf{c}(i,l,\phi) - \mathbf{c}(i,l,\phi-1)\big)\times \\ \big(\mathbf{c}(i,l,\phi) - \mathbf{c}(i,l,\phi+1)\big) > 0\end{array} \quad (|\phi| \leq 89), \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$\sigma(i) = 4.5 - (i-1)\frac{3.75}{I-1}, \tag{10}$$

$\{i \in \mathbb{N} : 1 \leq i \leq I\}$, I is the number of channels (32), * denotes convolution, and $\mathbb{N}$ is the set of natural numbers.

The skeleton cross-correlograms are subsequently pooled across frequency and time thus:

$$\bar{\mathbf{s}}(\phi) = \sum_{i,l} \mathbf{s}(i, l, \phi) \tag{11}$$

This pooled skeleton cross-correlogram is used to obtain "global" estimates of the target signal and interferer azimuths (ϕ_t and ϕ_i respectively), which are identified using the following procedure:

$$\phi_t = \min(\phi_1, \phi_2) \tag{12}$$

$$\phi_i = \max(\phi_1, \phi_2) \tag{13}$$

where

$$\phi_1 = \arg\max_{\psi_\phi} \bar{\mathbf{s}}(\psi_\phi), \tag{14}$$

$$\phi_2 = \arg\max_{\psi_\phi} \{\bar{\mathbf{s}}(\psi_\phi) : \phi_1 \notin \psi_\phi\} \tag{15}$$

and $\{\psi_\phi \in \phi : (\bar{\mathbf{s}}(\phi) - \bar{\mathbf{s}}(\phi - 1))(\bar{\mathbf{s}}(\phi) - \bar{\mathbf{s}}(\phi + 1)) > 0\}$. Note that the target is consistently placed on the left and thus the azimuths are assigned accordingly.

1.1.4 Binary Mask

The azimuthal cross-correlograms are used to calculate the binary time–frequency (T–F) mask **m** by making "local" estimates of the relative strength of the target and interfering signals at the obtained global azimuths thus:

$$\mathbf{m}(i, l) = \begin{cases} 1 & \text{if } \mathbf{c}(i, l, \phi_t) > \mathbf{c}(i, l, \phi_i) \\ & \text{and } 10\log_{10}\left(\dfrac{\mathbf{c}(i, l, \phi_t)}{\mathring{\mathbf{c}}}\right) > \Theta_c \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

where

$$\mathring{\mathbf{c}} = \max_{i,l,\phi} \mathbf{c}(i, l, \phi) \tag{17}$$

Generally Θ_c was set to −160 dB.

Once the binary mask has been estimated, two additional checks are performed on the mask: an interaural level difference (ILD) check for azimuthal estimate consistency and a rate threshold check. These are described below.

1.1.5 Interaural Level Difference

The ILD value for each T–F unit in frequency channels above 2.8 kHz (denoted v) that has a corresponding mask value of one is checked against an ILD template ζ to ensure azimuthal estimate consistency. The ILD template specifies the ILD at each available angle and in each frequency channel v; it was calculated using pseudo-anechoic HRTFs (see Section 3.4) and white noise. A zero is written to the mask if the ILD value deviates from the template by more than 1 dB:

$$\mathbf{m}(v, l) = \begin{cases} 0 & \text{if } |\text{ILD}(v, l) - \zeta(v, \phi_t)| > 1 \text{ dB} \\ \mathbf{m}(v, l) & \text{otherwise} \end{cases} \tag{18}$$

where

$$\text{ILD}(i, l) = 10\log_{10}\left(\frac{\acute{\mathbf{u}}_{\text{L}}(i, l)}{\acute{\mathbf{u}}_{\text{R}}(i, l)}\right), \tag{19}$$

$$\acute{\mathbf{u}}_k(i, l) = \left(\mathbf{u}_k(i, l)^{3.333}\right)^2 \tag{20}$$

1.1.6 Rate Threshold

Energy values where the corresponding mask value is one are compared to a running energy average Ξ, calculated in each frequency channel over a 200 ms (20 frame) window with 100 ms (10 frame) overlap. If the ratio of these values exceeds a rate threshold then a zero is written to the mask thus:

$$\mathbf{m}(i, l) = \begin{cases} 0 & \text{if } 10\log_{10}\left(\dfrac{\acute{\mathbf{u}}_{\text{LR}}(i, l)}{\Xi(i, l)}\right) > \Theta_r \\ \mathbf{m}(i, l) & \text{otherwise} \end{cases} \tag{21}$$

where

$$\mathbf{u}_{\text{LR}} = \left(\frac{1}{2}\left(\mathbf{u}_{\text{L}}^{3.333}(i, l) + \mathbf{u}_{\text{R}}^{3.333}(i, l)\right)\right)^{0.3}, \tag{22}$$

$\acute{\mathbf{u}}_{\text{LR}}$ was calculated as in (20) and Θ_r is the rate threshold set to −11 dB. This check was introduced in the original model [4] because it was found to be effective in regions with a low signal-to-noise ratio. In these regions, where target energy is weak, azimuth estimation is likely to be inaccurate. Hence, mask estimation may be more accurate if these T–F regions with low energy are rejected.

1.1.7 Resynthesis

Once the binary mask has been calculated, the output can be resynthesized. However, the evaluation procedure described later does not require a resynthesized output, and so this is not implemented. Furthermore, in the original model a spectral energy normalization procedure was introduced in order to undo the spectral envelope distortion caused by reverberation. This procedure was applied at resynthesis and, hence, also not implemented here.

1.2 Summary

This section presented a separation algorithm that estimates the relative strength of two competing signals arising from spatially-separate sound sources and separates them by calculating a binary mask. The algorithm includes a precedence model to suppress information following an onset. For comparison, the precedence model can be bypassed by setting $G = 0$ in (4). The precedence model will also be compared with models presented in the following section.

2 REPLACING THE PRECEDENCE MODEL

This section describes the incorporation of four alternative computational precedence models into the baseline separation algorithm. In order to attempt to improve the performance of the baseline separation algorithm, each of a selection of the numerous computational precedence and binaural localization models proposed in the literature was

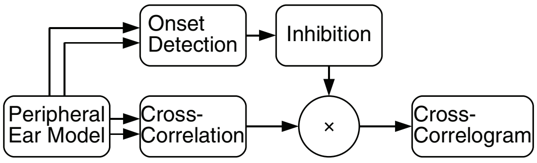

Fig. 3. Martin's [10] computational implementation of Zurek's [13] precedence model.

incorporated into the algorithm by replacing the baseline precedence model. Models proposed by Martin [10] (Section 2.1), Faller & Merimaa [7] (Section 2.2), Lindemann [8,16] (Section 2.3), and Macpherson [9] (Section 2.4) are presented. In each case, the baseline separation algorithm is retained, but the precedence model—and in some cases the peripheral ear model—is replaced by that of the model under test. As mentioned in the previous section, the precedence model takes the output of the peripheral ear model and returns the cross-correlogram for each frame and frequency channel $\mathbf{c}(i, l, \tau)$; the separation algorithm then warps this to the azimuthal domain and uses it to calculate the binary mask as in (16).

It should be noted that this study is designed to test the performance of the combination of the baseline algorithm and the computational precedence models. No judgments are or will be made about the technical quality, biological plausibility or even the localization accuracy of the models, although clearly the latter will have a significant influence on the separation performance.

Although it has long been known that ILD plays a significant role in localization, especially at high frequencies [17], in this work ILD is used only to check the consistency of azimuth estimations for frequencies above 2.8 kHz (see Section 1.1). It is not incorporated into the precedence modeling because the original precedence models do not incorporate it, nor do they provide an obvious mechanism for its incorporation.

2.1 Martin's Model

Martin's [10] model (shown in Fig. 3) is an implementation of Zurek's [13] account of precedence. The lower path of the model performs localization using ITD. The upper path of the model takes effect when sharp onsets are present in the signal. When such an onset is detected, a brief period of inhibition is triggered that suppresses the contribution of the lower path. The inhibition is maximal 1.5 ms after the onset and recovers over approximately 10 ms.

Unfortunately, Martin's paper lacks some crucial details necessary to implement the model accurately. Specifically, Martin's paper lacks details regarding the filter to calculate the "excitation envelope" and about the numeric levels of the numerous signals that are calculated. However, there is only one conceptual difference between the baseline precedence model and Martin's model: the point at which the inhibition is applied. In the baseline model, inhibition is applied to the fine structure before it is cross-correlated, whereas in Martin's model inhibition is applied to the running cross-correlation. Consequently, the implementation of Martin's model in the current study is heavily based upon the baseline precedence model.

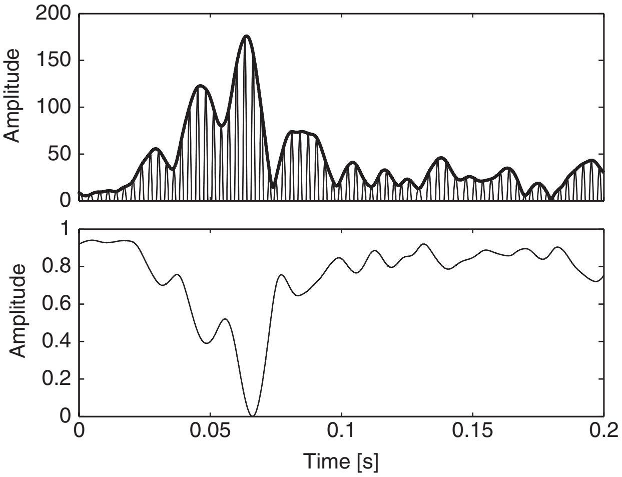

Fig. 4. Example of the processing in Martin's precedence model using the same signal as in Fig. 2. *Top:* Half-wave rectified gammatone filter output. *Bottom:* The resulting inhibitory signal.

In the implementation, first the "excitation envelope" $\mathbf{x}$ is calculated from the Hilbert envelope thus:

$$\mathbf{x}_k(i, n) = \varepsilon_k(i, n) * h_{\mathrm{lp}}(n) \tag{23}$$

where h_{lp} was given in (3), except that in this case the time constant $\alpha_p = \alpha_m = 1.5$ ms. Following this, a mono excitation envelope $\mathbf{x}_{LR}$ is calculated:

$$\mathbf{x}_{\mathrm{LR}}(i, n) = \frac{1}{2}\big(\mathbf{x}_{\mathrm{L}}(i, n) + \mathbf{x}_{\mathrm{R}}(i, n)\big) \tag{24}$$

and subsequently normalized independently for each frequency channel to be in the range [0,1]. The inhibitory signal ι is calculated from this excitation envelope thus:

$$\iota(i, n) = \max\Big(1 - \big(G \cdot \mathbf{x}_{\mathrm{LR}}(i, n)\big), 0\Big) \tag{25}$$

The inhibited running cross-correlation $\mathbf{c}_\iota$ is then calculated in the following way:

$$\mathbf{c}_\iota(i, n, \tau) = \iota(i, n)\,\acute{\mathbf{c}}(i, n, \tau) \tag{26}$$

where

$$\acute{\mathbf{c}}(i, n, \tau) = \mathbf{h}_{\mathrm{L}}\big(i, \max(n + \tau, n)\big)\mathbf{h}_{\mathrm{R}}\big(i, \max(n - \tau, n)\big) \tag{27}$$

Finally, these cross-correlations are averaged over a three-frame rectangular window to produce the cross-correlograms:

$$\mathbf{c}(i, l, \tau) = \frac{1}{3M}\sum_{d=1}^{3M} \mathbf{c}_\iota\big(i, (l-1)M + d, \tau\big) \tag{28}$$

As with the following models, subsequent processing of the cross-correlograms, grouping and separation routines is identical to that described in Section 1. An example of the inhibition procedure is shown in Fig. 4.

2.2 Faller & Merimaa's Model

The model proposed by Faller & Merimaa [7] differs from other computational precedence models by suggesting that some precedence effects can be modeled by calculating Interaural Coherence (IC). Specifically, if a dichotic signal is coherent then this is a good indication that the obtained ITD and ILD correspond to the sound's true direction. IC χ is calculated in each frequency band as the maximum value of the running normalized cross-correlation $\hat{\mathbf{c}}$:

$$\chi(i, n) = \max_{\tau} \hat{\mathbf{c}}(i, n, \tau) \tag{29}$$

This gives a result in the interval [0,1], with a value of one indicating that the signals are perfectly coherent and hence that the elicited cues are indicative of the sound's true direction. It is therefore necessary to specify a threshold for cue selection. According to Faller & Merimaa [7], this is a trade-off between selecting reliable cues that correspond closely to free-field conditions and maximizing the proportion of the input signals that contributes to localization. They also note that the optimal choice of threshold is likely to be dependent on the acoustic environment.

In terms of implementation, the first stage of the model is the peripheral auditory processing. Faller & Merimaa [7] suggest the use of a model of neural transduction proposed in [18]. This model recreates the compression and half-wave rectification that has been observed by numerous researchers in auditory physiology but does not enhance onsets. The employed process is summarized as follows:

- Each Hilbert envelope output of the gammatone filterbank ε_k is compressed by raising it to the power 0.23 and then squared;
- This envelope is then filtered with a fourth-order FIR low-pass filter with a cut-off frequency of 425 Hz;
- The resulting envelopes $\acute{\varepsilon}_k$ are half-wave rectified and then re-combined with the half-wave rectified gammatone filterbank output thus:

$$\mathbf{h}_k(i, n) = \frac{\acute{\varepsilon}_k(i, n)}{\varepsilon_k(i, n)} \max\bigl(\gamma_k(i, n), 0\bigr) \tag{30}$$

where $\mathbf{h}_k$ is the modeled IHC response and γ_k is the gammatone filter fine structure.

The cross-correlograms are calculated using the IHC-modeled data. As stated above, this model requires the calculation of normalized running cross-correlation, which is of the form

$$\hat{\mathbf{c}}(i, n, \tau) = \frac{\acute{\mathbf{c}}(i, n, \tau)}{\sqrt{\mathbf{a}_{\mathrm{L}}(i, n, \tau)\mathbf{a}_{\mathrm{R}}(i, n, \tau)}} \tag{31}$$

where

$$\begin{aligned}\acute{\mathbf{c}}(i, n, \tau) = {} & \frac{1}{\alpha_f}\mathbf{h}_{\mathrm{L}}\bigl(i, \max(n+\tau, n)\bigr)\mathbf{h}_{\mathrm{R}}\bigl(i, \max(n-\tau, n)\bigr) \\ & + \left(1 - \frac{1}{\alpha_f}\right)\acute{\mathbf{c}}(i, n-1, \tau),\end{aligned} \tag{32}$$

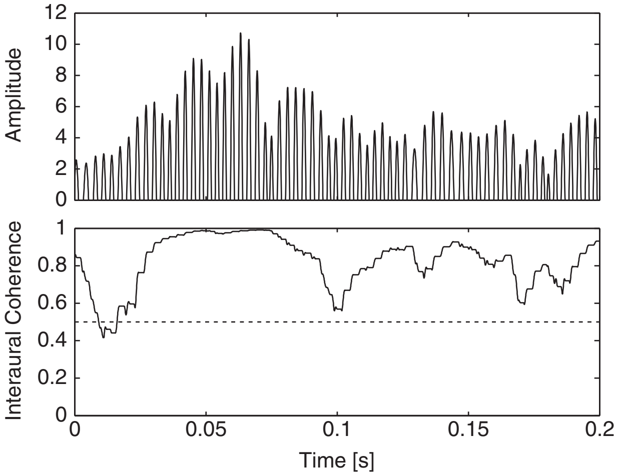

Fig. 5. Example of the processing in Faller & Merimaa's model using the same signal as in Fig. 2. *Top:* The IHC-modeled data. *Bottom:* The IC signal. Dashed line shows the IC threshold. Grey regions do not contribute to localization.

$$\begin{aligned}\mathbf{a}_{\mathrm{L}}(i, n, \tau) = {} & \frac{1}{\alpha_f}\mathbf{h}_{\mathrm{L}}^2\bigl(i, \max(n+\tau, n)\bigr) \\ & + \left(1 - \frac{1}{\alpha_f}\right)\mathbf{a}_{\mathrm{L}}(i, n-1, \tau),\end{aligned} \tag{33}$$

$$\begin{aligned}\mathbf{a}_{\mathrm{R}}(i, n, \tau) = {} & \frac{1}{\alpha_f}\mathbf{h}_{\mathrm{R}}^2\bigl(i, \max(n-\tau, n)\bigr) \\ & + \left(1 - \frac{1}{\alpha_f}\right)\mathbf{a}_{\mathrm{R}}(i, n-1, \tau),\end{aligned} \tag{34}$$

and α_f is the time constant of the exponentially decaying window, chosen to be the number of samples corresponding to 10 ms. The cross-correlograms are calculated by averaging only the running normalized cross-correlations within a given frame for which the corresponding IC value χ exceeds a threshold value Θ_χ:

$$\mathbf{c}(i, l, \tau) = \begin{cases} 0 & \text{if } \Psi = \varnothing \\ \dfrac{1}{|\Psi|}\displaystyle\sum_{d \in \Psi} \hat{\mathbf{c}}(i, d, \tau) & \text{otherwise} \end{cases} \tag{35}$$

where $\{\Psi \in n\colon (l-1)M + 1 \leq n \leq lM,\ \chi(i, n) \geq \Theta_\chi\}$, χ was given in (29), $\varnothing$ is the empty set, and Θ_χ is chosen to be 0.5, corresponding to two simultaneous and coherent onsets arising from two statistically-independent sound sources. An example of this processing is shown in Fig. 5.

2.3 Lindemann's Model

Lindemann's [8,16] model can be considered as an extension of Jeffress's [19] original cross-correlation theory of sound localization. The model is extended with two components: "monaural detectors" and a "contralateral-inhibition mechanism" (an inhibition along the τ-axis). This inhibition is achieved through two components: a static inhibition component and a dynamic inhibition component, the latter of which is intended to simulate the precedence

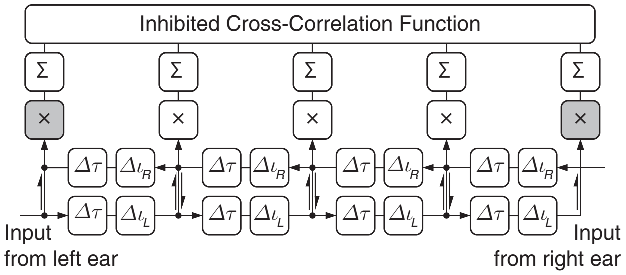

Fig. 6. The architecture of Lindemann's binaural localization model [8]. Adapted from [6,8].

effect. Although intended for stationary signals, the cross-correlation-based architecture lends itself well to this application. However, the suitability of the model to non-stationary signals remains unclear.

The architecture of the localization model is summarized in Fig. 6. The inhibition is derived from the contralateral signals and also from previous calculations of the cross-correlation. Furthermore, the inhibition is triggered by peaks in the primary cross-correlation and decays with a time constant of 10 ms. Additionally, monaural detectors (indicated by the grey multiplication boxes at the beginning of each delay line in Fig. 6) are included in order to lateralize the input even if only one ear signal is present and cross-correlation fails.

In terms of implementation, the peripheral auditory processing of the baseline algorithm is retained since Lindemann states that the exact nature of the peripheral processing is inconsequential to the operation of the model. According to Lindemann, the first step is to normalize the binaural signals to have a maximum value of one. However, the input level is critical to the model's operation; this is discussed toward the end of the section (see "The Operating Point"). Following the normalization, the modified inputs to the model, $\acute{\mathbf{h}}_{\mathrm{L}}$ and $\acute{\mathbf{h}}_{\mathrm{R}}$, are defined thus:

$$\acute{\mathbf{h}}_{\mathrm{L}}(i,n+1,\tau+1) = \begin{cases} \acute{\mathbf{h}}_{\mathrm{L}}(i,n,\tau)\iota_{\mathrm{L}}(i,n,\tau) & -T \leq \tau \leq T-1 \\ \mathbf{h}_{\mathrm{L}}(i,n+\tau) & \tau = T \end{cases} \tag{36}$$

$$\acute{\mathbf{h}}_{\mathrm{R}}(i,n+1,\tau-1) = \begin{cases} \acute{\mathbf{h}}_{\mathrm{R}}(i,n,\tau)\iota_{\mathrm{R}}(i,n,\tau) & -T+1 \leq \tau \leq T \\ \mathbf{h}_{\mathrm{R}}(i,n+\tau) & \tau = -T \end{cases} \tag{37}$$

where T is the maximum lag in samples. Note here that the outputs of the peripheral processor $\mathbf{h}_{\mathrm{L}}$ and $\mathbf{h}_{\mathrm{R}}$ have had zeros placed between alternate samples in order to halve the sample period. The inhibitory components ι_{L} and ι_{R} are derived from the contralateral signal in the following way:

$$\iota_{\mathrm{L}}(i,n,\tau) = \left(1-\acute{\mathbf{h}}_{\mathrm{R}}(i,n,\tau)\right)\left(1-\Phi(i,n-1,\tau)\right) \tag{38}$$

$$\iota_{\mathrm{R}}(i,n,\tau) = \left(1-\acute{\mathbf{h}}_{\mathrm{L}}(i,n,\tau)\right)\left(1-\Phi(i,n-1,\tau)\right) \tag{39}$$

Here, Φ is the dynamic inhibitory component, which is derived from the cross-correlation product $\acute{\mathbf{c}}$ in the following way:

$$\Phi(i,n,\tau) = \acute{\mathbf{c}}(i,n-1,\tau) + \Phi(i,n-1,\tau)e^{-T_d/\alpha_{\mathrm{inh}}}\left(1-\acute{\mathbf{c}}(i,n-1,\tau)\right) \tag{40}$$

where T_d is half the sample period and α_{inh} is the fade-off time constant (10 ms). The running cross-correlation is calculated as follows:

$$\acute{\mathbf{c}}(i,n,\tau) = \Big(p(\tau)+\big(1-p(\tau)\big)\acute{\mathbf{h}}_{\mathrm{R}}(i,n,\tau)\Big) \times\Big(p(-\tau)+\big(1-p(-\tau)\big)\acute{\mathbf{h}}_{\mathrm{L}}(i,n,\tau)\Big) \tag{41}$$

where p is the monaural sensitivity function such that $p(\tau) = 0.035e^{-(T+\tau)/6}$. The inhibited cross-correlation $\mathbf{c}_\iota$ is calculated from the running cross-correlation using an exponential window thus:

$$\mathbf{c}_\iota(i,n,\tau) = \left(1-e^{-T_d/T_{\mathrm{int}}}\right)\acute{\mathbf{c}}(i,n,\tau) + e^{-T_d/T_{\mathrm{int}}}\mathbf{c}_\iota(i,n-1,\tau) \tag{42}$$

where T_{int} is the integration time constant (5 ms). The cross-correlograms are calculated by averaging the running cross-correlations over the frame:

$$\mathbf{c}(i,l,\tau) = \frac{1}{M}\sum_{d=1}^{M}\mathbf{c}_\iota(i,(l-1)M+d,\tau) \tag{43}$$

The Operating Point

One difficulty in Lindemann's paper is the discussion of the "operating point" or "inhibition parameter" (c_{inh}). The parameter appears to be crucial for controlling the amount of inhibition. Although Lindemann states how it is derived, he does not discuss how it is implemented. Specifically, Lindemann [8] states that:

'The operating point is described by the "inhibition parameter" c_{inh} that is derived from the input signal having the greater amplitude. For pure tones with the amplitudes A_r (right input signal) and A_l (left input signal) the inhibition parameter is

$$c_{\mathrm{inh}} = \max\{A_r, A_l\} \qquad \text{with } 0 \leq c_{\mathrm{inh}} \leq 1$$

For stationary noise signals c_{inh} was derived analogously, A_r and A_l being the root-mean-square (before half-wave rectification), multiplied by $\sqrt{2}$. The noise signals were clipped after the half-wave rectification to avoid input signals greater than one.'

Clearly, although the inhibition parameter is "derived," there must be a mechanism that aims to achieve a given inhibition parameter (c_{inh}) at the input to the model. Consequently, here the input to the model $\mathbf{h}$ is derived in the following way, based on the above description and a target inhibition parameter c_{inh}:

$$\mathbf{h}_k(i,n) = \min\left(\max\left(\frac{c_{\mathrm{inh}}}{c_\gamma(i)}\gamma_k(i,n),0\right),1\right) \tag{44}$$

where

$$c_\gamma(i) = \max_k\sqrt{\frac{2}{N}\sum_{n=1}^{N}\gamma_k^2(i,n)}, \tag{45}$$

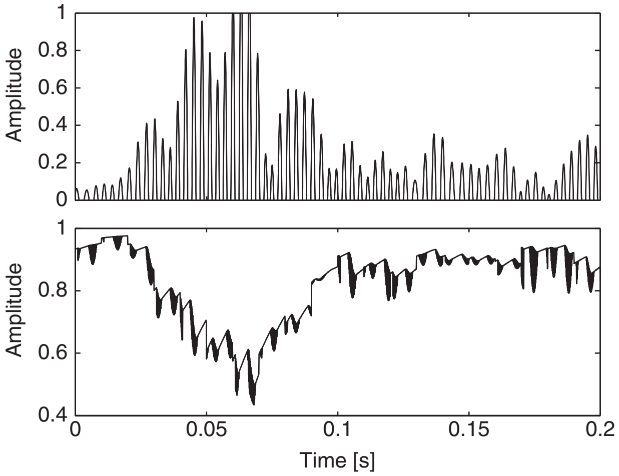

Fig. 7. Examples of the processing in Lindemann's model using the same signal as in Fig. 2. *Top:* The output of the peripheral processor. *Bottom:* The left inhibitory component ι_{L} with $\tau = 0$.

and N is the number of samples at the input. Lindemann states that the optimal value for $c_{\mathrm{inh}} = 0.3$ and hence this value is employed in the investigation. An example of the processing in the model is shown in Fig. 7.

2.4 Macpherson's Model

Macpherson [9] proposes a model for stereo imaging measurement. However, since the model is based on cross-correlation, it can be easily adapted for use in this work. The first stage of the model is the peripheral processing, however, there is insufficient information to accurately recreate this stage. Since this stage aims to recreate both the cochlear filtering and the half-wave rectification, adaptation and phase- and envelope-locking seen in auditory nerve responses, a combination of a gammatone filterbank and a Meddis IHC model are utilized in the peripheral processing.

The precedence modeling is introduced through the selection of "analysis points." Macpherson argues that performing a running cross-correlation for the entire signal length is inefficient. Therefore, a set of analysis points (samples) Ψ are chosen where local peaks occur across the left and right ear signals within the cross-correlation window M_c (2 ms, in samples) such that:

$$\Psi = \Psi_{\mathrm{L}} \cap \Psi_{\mathrm{R}} \tag{46}$$

where

$$\begin{aligned}\Psi_{\mathrm{L}} = \big\{n : \big(\mathbf{h}_{\mathrm{L}}(i,n) - \mathbf{h}_{\mathrm{L}}(i,n-1)\big)\\ \times\big(\mathbf{h}_{\mathrm{L}}(i,n) - \mathbf{h}_{\mathrm{L}}(i,n+1)\big) > 0\big\},\end{aligned} \tag{47}$$

$$\begin{aligned}\Psi_{\mathrm{R}} = \Big\{n + \mu : (\mathbf{h}_{\mathrm{R}}(i,n) - \mathbf{h}_{\mathrm{R}}(i,n-1)\big)\\ \times\big(\mathbf{h}_{\mathrm{R}}(i,n) - \mathbf{h}_{\mathrm{R}}(i,n+1)\big) > 0,\\ \mu \in \mathbb{Z}, \frac{-M_c}{2} \le \mu \le \frac{M_c}{2}, \mu \ne 0\Big\}\end{aligned} \tag{48}$$

At high frequencies, even with the envelope-locking characteristics of the IHC model, peaks can occur very close together, creating significant overlap of the cross-correlation windows. To reduce this inefficiency, the input is divided into frames of length $M_c/2$ and only the last analysis point from each frame is selected.

The cross-correlation $\acute{\mathbf{c}}$ is calculated for each member of Ψ with the peak at the center of the cross-correlation window. To simulate the precedence effect, an inhibited cross-correlation is calculated as a weighted average of cross-correlations that fall within the inhibition window 20 ms in length (two frames, in samples) after the initial analysis point. Unfortunately, Macpherson does not specify this weighting function, only stating that peaks that occur within 1–6 ms are suppressed. Consequently, the weighting window proposed in [10] is adapted and utilized here and the inhibited cross-correlation is calculated in the following way:

$$\mathbf{c}_{\iota}(i,n,\tau) = \begin{cases} 0 & \text{if } \psi = \varnothing \\ \dfrac{1}{|\psi|}\displaystyle\sum_{d\in\psi} w_m(x-n)\acute{\mathbf{c}}(i,d,\tau) & \text{otherwise} \end{cases}, \quad (n \in \Psi) \tag{49}$$

where $\{\psi \subset \Psi : n \le \Psi \le n + 2L\}$,

$$\acute{\mathbf{c}}(i,n,\tau) = \frac{1}{M_c+1}\sum_{d=n-\frac{M_c}{2}}^{n+\frac{M_c}{2}} \begin{pmatrix} \mathbf{h}_{\mathrm{L}}\big(i, \max(d+\tau, d)\big)\times \\ \mathbf{h}_{\mathrm{R}}\big(i, \max(d-\tau, d)\big) \end{pmatrix}, \quad (n \in \Psi), \tag{50}$$

$$w_m(n) = A\max\left(1 - G\frac{e}{\alpha_m}h_{\mathrm{lp}}(n), 0\right), \tag{51}$$

h_{lp} was as in Martin's model (see Section 2.1), α_m was defined in Martin's model (set in samples to 1.5 ms), G is the inhibitory gain (set to 1), and A is set to give unity gain at DC. Last, these weighted cross-correlations are averaged across the duration of the frame to form the cross-correlograms thus:

$$\mathbf{c}(i,l,\tau) = \begin{cases} 0 & \text{if } \acute{\psi} = \varnothing \\ \dfrac{1}{|\acute{\psi}|}\displaystyle\sum_{d\in\acute{\psi}} \mathbf{c}_{\iota}(i,d,\tau) & \text{otherwise} \end{cases} \tag{52}$$

where $\{\acute{\psi} \subset \Psi : (l-1)M + 1 \le \Psi \le lM\}$. An example of this processing is shown in Fig. 8.

2.5 Summary

This section has presented precedence models suggested by Martin [10], Faller & Merimaa [7], Lindemann [8,16], and Macpherson [9]. A summary of the peripheral processing and precedence processing in each model is provided in Table 1.

3 EXPERIMENTAL PROCEDURE

This section describes the procedure used to test the models and includes specific discussions of independent

Table 1. Summary of peripheral ear and precedence processing for each model.

Model	Peripheral Ear Model	Precedence Model Outline
Baseline	Half-wave rectifier	Calculates an inhibitory signal from onset information, uses this to suppress fine structure output by the peripheral ear model that may be corrupted by reverberation
Martin	Half-wave rectifier	Calculates an inhibitory signal from onset information, uses this to suppress localization information that may be corrupted by reverberation
Faller & Merimaa	Bernstein et al. [18]	Localizes sounds using cues that exceed an interaural coherence (IC) threshold
Lindemann	Clipper and half-wave rectifier	Extends Jeffress's [19] cross-correlation theory of sound localization with several components, including a dynamic inhibition component intended to simulate the precedence effect.
Macpherson	Meddis et al. [20]	Performs localization exclusively at local peaks and weights them in a manner similar to that employed in Martin's model

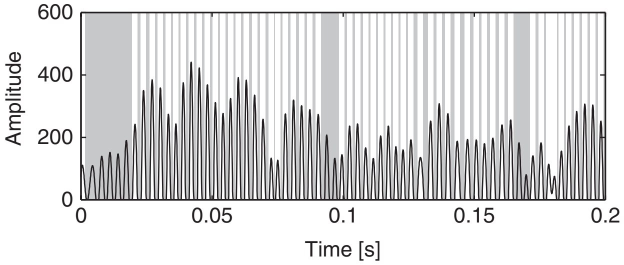

Fig. 8. Example of the processing in Macpherson's precedence model using the same signal as in Fig. 2. The cross-correlation windows are shown in white. Grey regions do not contribute to localization.

variables, the choice of metric, signals, and how the Binaural Room Impulse Responses (BRIRs) were obtained.

3.1 Independent Variables

The models were tested in a range of mixture conditions similar to those tested in [4]. A range of conditions was employed to ensure that the performances (reported later) were representative of a range of realistic conditions offering a varying degree of difficulty. However, only RT_{60} will be compared in the results, with model performances reported as means calculated across the other variables. Specifically, the models were tested under the following conditions:

- Target/interferer azimuthal separations of 10°, 20°, and 40° (i.e., ±5°, ±10°, and ±20° with respect to the frontal median plane), with the target on the left;
- Target-to-Interferer Ratios (TIRs) of 0, 10, and 20 dB (RMS);
- The following interferers: white noise, male speech, and a modern piece of rock music (see Section 3.3);
- Four real rooms and an anechoic room, selected to have a range of RT_{60}s, but also with a range of other acoustic properties as shown in Table 2 and discussed in Section 3.4.

These variables give rise to 135 experimental combinations.

Table 2. Room acoustical properties, including RT_{60}, direct-to-reverberant ratio (DRR), and initial time delay gap (ITDG).

Room	RT_{60} [s]	DRR [dB]	ITDG [ms]
A	0.32	8.72	6.09
B	0.47	5.31	9.66
C	0.68	8.82	11.9
D	0.89	6.12	21.6

3.2 Choice of Metric

Many applications of this work are likely to entail further machine processing rather than human listening (e.g., automatic speech recognition). Since human perception is not necessarily an accurate predictor of machine performance in such applications, a metric that objectively assesses the attained degree of separation is deemed more appropriate than a listening-based assessment.

According to Li and Wang [21] a widely utilized objective metric for assessing source separation is Signal-to-Noise Ratio:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_n s^2(n)}{\sum_n\left(\hat{s}(n) - s(n)\right)^2}\right) \tag{53}$$

where s is the target signal and $\hat{s}$ is the estimated target signal. Note that the denominator is a summation of a difference signal and thus incorporates any and all differences between the target and estimated target. In this study this includes the reverberation present in the mixture. The reverberation contributes differently to the target and estimated target, increasing the magnitude of the denominator and lowering the SNR. Furthermore, the calculated SNR is likely to vary dramatically according to the nature of the reverberation. Hence, for the same signals and binary mask, SNR is likely to demonstrate large inconsistencies between different acoustic environments. This prevents meaningful comparison of separation algorithms across different acoustic conditions. Consequently a novel metric is proposed, loosely based on binary mask error [22] or labeling accuracy [23], to assess the separation performance of the algorithm. The metric—the Ideal Binary Mask Ratio

(IBMR) [24]—assesses the calculated mask **m** by comparing it directly with the Ideal Binary Mask $\mathbf{m}_{\text{IBM}}$ thus:

$$\text{IBMR} = \frac{\lambda}{\lambda + \rho} \tag{54}$$

where λ denotes the number of target T–F units the two masks have in common and ρ denotes the number of T–F units that differ between the masks. These counts are calculated thus:

$$\lambda = \sum_{i,l} \mathbf{m}(i,l) \wedge \mathbf{m}_{\text{IBM}}(i,l), \tag{55}$$

$$\rho = \sum_{i,l} \mathbf{m}(i,l) \oplus \mathbf{m}_{\text{IBM}}(i,l) \tag{56}$$

where ∧ denotes binary logical AND and ⊕ denotes binary logical XOR. The IBM is calculated using the following logic:

$$\mathbf{m}_{\text{IBM}}(i,l) = \begin{cases} 1 & \text{if } 10\log_{10}\left(\dfrac{\acute{\mathbf{u}}_t(i,l)}{\acute{\mathbf{u}}_i(i,l)}\right) > \Theta_{\text{IBM}} \\ 0 & \text{otherwise} \end{cases} \tag{57}$$

where $\acute{\mathbf{u}}_t$ and $\acute{\mathbf{u}}_i$ denote the clean target and interferer energy respectively and Θ_{IBM} is a threshold value set to 0 dB. See Section 1 (20) for the calculation of $\acute{\mathbf{u}}$.

3.3 Signals

As stated above, similar interferers to those used in [4] were used in the experimental procedure. The target signal was a four-second excerpt of female speech taken from the European Broadcasting Union Sound Quality Assessment Material [25], chosen because many applications are likely to be based on speech. The interfering signals were chosen to be representative of signals encountered in the real world and to provide a range of challenges. They were: a rock music track (“Action!” by Razorlight), white noise, and an excerpt of male speech also taken from [25]. The speech segments were chosen to incorporate a wide range of phonemes. The white noise, although perhaps unnatural, provides a slightly different challenge compared to the speech and music: the white noise has more energy at high frequencies than the other signals (and less at lower frequencies) and so masking is likely to occur at higher frequencies than it would for the other interferers.

3.4 Binaural Room Impulse Responses

It was decided to use Binaural Room Impulse Responses (BRIRs) captured in real rooms rather than simulating them due to the generally poor subjective quality of responses calculated using acoustic models. The responses were captured in a variety of rooms at the University of Surrey using a Cortex (mk.2) Head and Torso Simulator (HATS) and Genelec 8020A loudspeaker. The loudspeaker was 1.5 m from the HATS in order to maintain commonality with the algorithm on which this research is based [4]. The loudspeaker replayed sine sweeps that were deconvolved to produce the impulse responses. Acoustical properties for each room are shown in Table 2. Measurements of RT_{60} were obtained according to [26] using an interrupted pink noise method with six microphone positions and two loudspeaker positions (12 measurements in total). In accordance with the standard, the overall room RT_{60} is calculated by averaging the 500 Hz and 1 kHz bands. For the anechoic condition, a similar procedure was used and impulse responses were obtained using a pseudo-anechoic approach whereby the responses were simply truncated before the first reflection, having been captured in a large room.

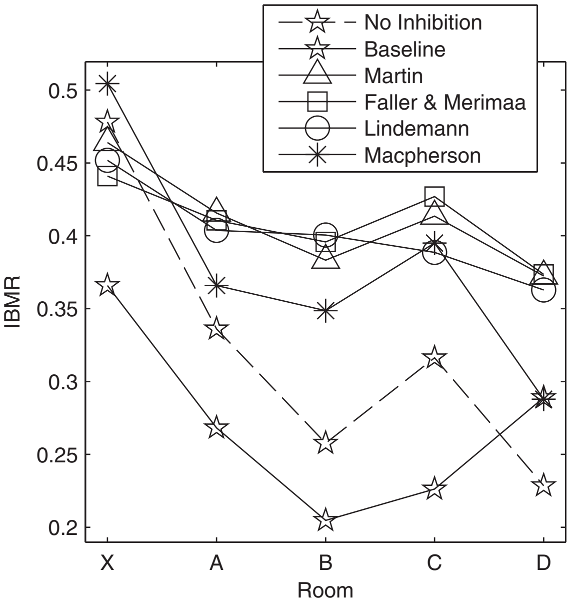

Fig. 9. Mean model performances showing IBMR versus room.

3.5 Summary

The following independent variables were used to test the models:

- Three target/interferer azimuthal separations;
- Three Target-to-Interferer Ratios (TIRs);
- Three interfering signals;
- Five acoustic environments with different RT_{60}s.

The performance metric was the Ideal Binary Mask Ratio (IBMR).

4 RESULTS AND DISCUSSION

The results from the study are given in Fig. 9. The plot shows IBMR versus room with the data averaged over all experimental conditions. The data are compared to “No Inhibition,” i.e., the data obtained from the baseline algorithm, except that the precedence model is bypassed by setting $G = 0$ (see (4)). Plotting the data obtained without precedence processing demonstrates the performance gain achieved by each of the precedence models.

There are five points to note about the plot:

1. The uninhibited model performs well for the anechoic condition, although the performance drops rapidly with the RT_{60} of the room.

2. The baseline model performs poorly and is outperformed by the uninhibited model until the room with the longest RT_{60}.
3. Martin's model appears to perform well in all conditions.
4. The models of Faller & Merimaa and Lindemann demonstrate average performance for the anechoic condition but perform favorably for the other acoustic conditions.
5. Macpherson's model performs well in the anechoic condition but performs less favorably in rooms with longer RT_{60}s compared to most other models.

From these results, three observations can be made about the data. First, the uninhibited model performs better than many of the precedence models for the anechoic condition. However, the performance drops off rapidly and many of the precedence models out-perform the uninhibited model for subsequent reverberant rooms. This may be because any precedence processing removes information that may be corrupted by reverberation. However, when no reverberation is present, this strategy removes information that would otherwise contribute to localization and hence to source identification. As RT_{60} increases the amount of usable localization information decreases and so the precedence models begin to out-perform the uninhibited model.

Second, the baseline precedence model appears to provide no performance gain until it is tested in rooms with longer RT_{60}s. As with some other models this is because, at shorter RT_{60}s, the model is excessively removing information that would otherwise positively contribute to localization and separation. It is not until longer RT_{60}s that this strategy becomes beneficial. This suggests that in order to optimize the performance of the separation algorithm, the precedence model should adapt its processing to the acoustic conditions. For example, the inhibitory gain factor G in the baseline and Martin models (see (4)) or the IC threshold Θ_χ in Faller & Merimaa's model (see (35)) may need to increase as the acoustic conditions deteriorate. Recent work, such as that described in [27], has demonstrated how the perceptual salience of reverberation changes as a function of both room acoustics and signal characteristics. Such work could make an important contribution to an adaptive precedence model for use in the baseline algorithm.

Last, in rooms with medium to long RT_{60}s, many of the models perform comparably (within about 0.05 IBMR). However, the models of Martin and Faller & Merimaa generally perform best across the rooms. Again, the relatively poorer performance in rooms with short RT_{60}s may be due to the removal of information that would otherwise positively contribute to localization.

5 CONCLUSIONS

The aim of this paper was to investigate whether an enhanced precedence model can improve the separation performance of a baseline separation algorithm. The results above have shown that an enhanced precedence model can improve the separation performance of the baseline separation algorithm. Precedence models based on those proposed by Martin and by Faller & Merimaa showed consistently good performance across the rooms. It was noted earlier that Faller & Merimaa [7] state that setting the IC threshold in their model is a trade-off between selecting reliable cues that correspond closely to free-field conditions and maximizing the proportion of the input signals contributing to localization. The results shown in this paper reflect this and indicate that a dynamic component of the precedence models may be necessary in order to adapt the precedence processing to the acoustic conditions, thus maximizing the separation performance of the algorithm. This hypothesis has been supported in recent work that demonstrated that the precedence models can be optimized (in order to offer improved performance) for each room [28,29]. An interesting area for future work will be to build an algorithm that can automatically estimate these optimal precedence parameters by extracting the relevant acoustical parameters from the input.

6 ACKNOWLEDGMENTS

The authors wish to thank Joshua Brooks and Laurent Simon for their help in capturing the BRIRs; Dr. Richard O. Duda from San Jose State University, California, for his assistance with implementing the Lindemann model; and the reviewers of the initial draft of this paper, for their valuable feedback and advice. This work was supported by the EPSRC.

7 REFERENCES

[1] E. Blanco-Martin, F. J. Casajús-Quirós, J. J. Gómez-Alfageme, and L. I. Ortiz-Berenguer, "Objective Measurement of Sound Event Localization in Horizontal and Median Planes," *J. Audio Eng. Soc.*, vol. 59, pp. 124–136 (2011 Mar.).

[2] E. K. Kokkinis and J. Mourjopoulos, "Unmixing Acoustic Sources in Real Reverberant Environments for Close-Microphone Applications," *J. Audio Eng. Soc.*, vol. 58, pp. 907–922 (2010 Nov.).

[3] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The Precedence Effect," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1633–1654 (1999 Oct.).

[4] K. J. Palomäki, G. J. Brown, and D. Wang, "A Binaural Processor for Missing Data Speech Recognition in the Presence of Noise and Small-Room Reverberation," *Speech Comm.*, vol. 43, no. 4, pp. 361–378 (2004 Sep.).

[5] H.-M. Park and R. Stern, "Missing Feature Speech Recognition Using Dereverberation and Echo Suppression in Reverberant Environments," *P. IEEE Int. Conf. Acoust. Speech Signal Proc.*, vol. 4, pp. 381–384 (2007).

[6] J. Braasch, "Modelling of Binaural Hearing," in *Comm. Acoust.*, J. Blauert, Ed., pp. 75–108 (Berlin Heidelberg : Springer-Verlag, 2005).

[7] C. Faller and J. Merimaa, "Source Localization in Complex Listening Situations: Selection of Binaural Cues Based on Interaural Coherence," *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3075–3089 (2004 Nov.).

[8] W. Lindemann, “Extension of a Binaural Cross-Correlation Model by Contralateral Inhibition. I. Simulation of Lateralization for Stationary Signals,” *J. Acoust. Soc. Am.*, vol. 80, no. 6, pp. 1608–1622 (1986 Dec.).

[9] E. A. Macpherson, “A Computer Model of Binaural Localization for Stereo Imaging Measurement,” *J. Audio Eng. Soc.*, vol. 39, pp. 604–622 (1991 Sep.).

[10] K. D. Martin, “Echo Suppression in a Computational Model of the Precedence Effect,” *IEEE Workshop Appl. Signal Proc. Audio Acoust.*, New Paltz, NY (1997).

[11] A. S. Bregman, *Auditory Scene Analysis* (Cambridge, MA : MIT Press, 1990).

[12] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, “An Efficient Auditory Filterbank Based on the Gammatone Function,” MRC Applied Psychology Unit, Cambridge, Technical Report (1987 Dec.).

[13] P. M. Zurek, “The Precedence Effect,” in *Dir. Hear.*, W. A. Yost and G. Gourevitch, Eds., pp. 85–105 (New York: Springer-Verlag, 1987).

[14] G. F. Kuhn, “Model for the Interaural Time Differences in the Azimuthal Plane,” *J. Acoust. Soc. Am.*, vol. 62, no. 1, pp. 157–167, (1977 Jul.).

[15] N. Roman, D. Wang, and G. J. Brown, “Speech Segregation Based on Sound Localization,” *J. Acoust. Soc. Am.*, vol. 114, no. 4, pp. 2236–2252 (2003 Oct.).

[16] W. Lindemann, “Extension of a Binaural Cross-Correlation Model by Contralateral Inhibition. II. The Law of the First Wave Front,” *J. Acoust. Soc. Am.*, vol. 80, no. 6, pp. 1623–1630 (1986 Dec.).

[17] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, revised ed. (Cambridge, MA : MIT Press, 1997).

[18] L. R. Bernstein, S. Van de Par, and C. Trahiotis, “The Normalized Interaural Correlation: Accounting for NoSπ Thresholds Obtained with Gaussian and ’Low-Noise’ Masking Noise,” *J. Acoust. Soc. Am.*, vol. 106, no. 2, pp. 870–876 (1999).

[19] L. Jeffress, “A Place Theory of Sound Localization,” *J. Comp. Psych.*, vol. 41, no. 1, pp. 35–39 (1948 Feb.).

[20] R. Meddis, M. J. Hewitt, and T. M. Shackleton, “Implementation Details of a Computation Model of the Inner Hair-Cell Auditory-Nerve Synapse,” *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1813–1816, (1990 Apr.).

[21] Y. Li and D. Wang, “On the Optimality of Ideal Binary Time–Frequency Masks,” *Speech Comm.*, vol. 51, no. 3, pp. 230–239, (2009 Mar.).

[22] N. Li and P. C. Loizou, “Factors Influencing Intelligibility of Ideal Binary-Masked Speech: Implications for Noise Reduction,” *J. Acoust. Soc. Am.*, vol. 123, no. 3, pp. 1673–1682, (2008 Mar.).

[23] J. Woodruff and D. Wang, “Sequential Organization of Speech in Reverberant Environments by Integrating Monaural Grouping and Binaural Localization,” *IEEE T. Audio Speech Lang. Proc.*, vol. 18, no. 7, pp. 1856–1866, (2010 Sep.).

[24] C. Hummersone, R. Mason, and T. Brookes, “Ideal Binary Mask Ratio: A Novel Metric for Assessing Binary-Mask-Based Sound Source Separation Algorithms,” *IEEE T. Audio Speech Lang. Proc.*, vol. 19, no. 7, pp. 2039–2045 (2011).

[25] EBU SQAM, “Sound Quality Assessment Material for Subjective Listening Tests,” European Broadcasting Union, Tech. 3253-E, 1988, http://tech.ebu.ch/publications/sqamcd.

[26] BS EN ISO 3382, “Measurement of the Reverberation Time of Rooms with Reference to Other Acoustical Parameters” (2000).

[27] T. Zarouchas and J. Mourjopoulos, “Perceptually Motivated Signal-Dependent Processing for Sound Reproduction in Reverberant Rooms,” *J. Audio Eng. Soc.*, vol. 59, pp. 187–200 (2011 Apr.).

[28] C. Hummersone, R. Mason, and T. Brookes, “Dynamic Precedence Effect Modeling for Source Separation in Reverberant Environments,” *IEEE T. Audio Speech Lang. Proc.*, vol. 18, no. 7, pp. 1867–1871 (2010).

[29] C. Hummersone, “A Psychoacoustic Engineering Approach to Machine Sound Source Separation in Reverberant Environments,” Ph.D. dissertation, University of Surrey (2011).

THE AUTHORS

Christopher Hummersone

Russell Mason

Tim Brookes

Christopher Hummersone received the B.Mus. degree in music and sound recording (Tonmeister), from the University of Surrey, Guildford, U.K., in 2007 and the Ph.D. degree from the University of Surrey, in 2011. His research project centered on modeling the precedence effect for sound source localization and separation. He is currently a lecturer at the Institute of Sound Recording, University of Surrey. His research interests focus on machine listening for automated localization and separation of sound sources and for the evaluation of audio quality.

●

Russell Mason received the B.Mus. degree in music and sound recording (Tonmeister) from the University of Surrey, Guildford, U.K., in 1998, and the Ph.D. degree in audio engineering and psychoacoustics from the University of Surrey in 2002. He is currently a senior lecturer in the Institute of Sound Recording, University of Surrey. His research interests are focused on psychoacoustic engineering and include subjective experimentation on auditory spatial perception and the development of computational models of binaural hearing.

●

Tim Brookes received the B.Sc. degree in mathematics and the M.Sc. and D.Phil. degrees in music technology from the University of York, York, U.K., in 1990, 1992, and 1997, respectively. He was employed as a software engineer, recording engineer, and research associate before joining, in 1997, the academic staff at the Institute of Sound Recording, University of Surrey, Guildford, U.K., where he is now a senior lecturer in audio and director of research. His teaching focuses on acoustics and psychoacoustics and his research is in psychoacoustic engineering: measuring, modeling, and exploiting the relationships between the physical characteristics of sound and its perception by human listeners.