Dynamic Precedence Effect Modelling for Source Separation in Reverberant Environments

Christopher Hummersone, Russell Mason and Tim Brookes

***Abstract*—Reverberation continues to present a major problem for sound source separation algorithms. However, humans demonstrate a remarkable robustness to reverberation and many psychophysical and perceptual mechanisms are well documented. The precedence effect is one of these mechanisms; it aids our ability to localise sounds in reverberation. Despite this, relatively little work has been done on incorporating the precedence effect into automated source separation. Furthermore, no work has been carried out on adapting a precedence model to the acoustic conditions under test and it is unclear whether such adaptation, analogous to the perceptual Clifton effect, is even necessary. Hence, this study tests a previously proposed binaural separation/precedence model in real rooms with a range of reverberant conditions. The precedence model inhibitory time constant and inhibitory gain are varied in each room in order to establish the necessity for adaptation to the acoustic conditions. The study concludes that adaptation is necessary and can yield significant gains in separation performance. Furthermore, it is shown that the Initial Time Delay Gap and the Direct-to-Reverberant Ratio are important factors when considering this adaptation.**

***Index Terms*—Source Separation, Reverberation, Precedence Model**

I. Introduction

Automated audio source separation remains an area of high research interest. Separation algorithms can have many applications, including front-end processing for missing data speech recognition, and enhancement of hearing prostheses and communication devices such as mobile phones. In many of these 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 that these algorithms rely on. However, numerous human psychophysical and perceptual mechanisms for suppressing the effects of reverberation are well documented. One such mechanism is the precedence effect.

The precedence effect (for a review see [1]) is described in the perceptual literature as being an important mechanism for enhancing our ability to localise sounds in reverberant environments. Often referred to as the "law of the first wave front", the precedence effect describes an auditory mechanism which is able to give greater perceptual weighting to the first wave fronts of a sound—the direct sound—compared to later wave fronts arriving as reflections from surrounding surfaces. However, relatively little work has been carried out on incorporating precedence processing into separation algorithms that utilise spatial cues. To date, work in this area has been based on that of Palomäki et al. [2] (see also [3]). However, as Palomäki et al. note, the precedence model they utilise is somewhat simplified and further work could be done in order to improve its localisation capabilities.

Numerous computational precedence models have been proposed in the literature (see for example [4]–[8]). However, only Faller and Merimaa [5] discuss the necessity for the algorithm to adapt to different acoustic conditions. Furthermore, they do not discuss the computational mechanism to achieve this adaptation nor which acoustical factors affect it. Conversely, it is well documented that in humans the precedence effect has a dynamic component—the Clifton effect—that adjusts to the acoustic conditions in which the listener is located [9], [10]. The necessity for a computational Clifton–like processor has not been formally validated. Hence, this paper details a study investigating the extent to which computational precedence models need to adapt to different acoustic conditions in order to optimise separation performance and identifies the acoustical parameters that affect this adaptation. The study uses an implementation of the aforementioned algorithm of Palomäki et al. [2] as the baseline separation algorithm. The inhibitory time constant and inhibitory gain are adjusted over a range of acoustic conditions with the aim of optimising the separation performance.

The following section summarises the baseline separation and precedence algorithms. The experimental procedure is presented in Section III and the results are presented and discussed in Seciton IV. Conclusions are drawn in Section V and plans for future work are presented in Section VI.

II. The Separation and Precedence Algorithm

The separation algorithm utilised in this investigation is heavily based upon the aforementioned work described in [2] (note: although every attempt has been made to follow the principles of this model, due to implementation issues and modifications required to enable the evaluation method described below, the processing utilised is not identical). This section describes the pertinent aspects of the algorithm; the interested reader is referred to [11] for a detailed description of the implementation. The architecture of the experimental algorithm is summarised in Fig. 1.

The algorithm attempts to estimate the relative strength of two spatially–separate competing sound signals. This is achieved by cross-correlating the output of a peripheral ear model (a gammatone filterbank and half-wave rectifier) to obtain the azimuths of the sounds. Their relative strengths are estimated from the magnitude of the cross-correlation function at these azimuths. A precedence model is introduced to inhibit the fine structure before cross-correlation. The precedence model is based on the popular paradigm suggested by Zurek [12], which is the basis for several similar computational precedence models [2]–[4], [13]. The cross-correlation C is calculated in frequency channel i, time frame j and discrete lag τ (which represents Interaural Time Difference (ITD)) thus:

$$C(i,j,\tau) = \sum_{x=0}^{3M-\tau-1} r_L\big(i,(j-1)M+x+\tau\big)\,r_R\big(i,(j-1)M+x\big) \quad (1)$$

where M is the frame length in samples (10 ms) and r is the precedence–modelled fine structure, which is calculated for ear k thus:

$$r_k(i,n) = \max\Big(h_k(i,n) - G\big(h_{lp}(n) * \varepsilon_k(i,n)\big), 0\Big), \quad (2)$$

where h_k is the half-wave rectified output of the gammatone filterbank at sample index n, ε_k is the Hilbert envelope output of the gammatone filterbank, G is an inhibitory gain factor, $*$ denotes convolution and h_{lp} is a low-pass onset-de-emphasising filter such that

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

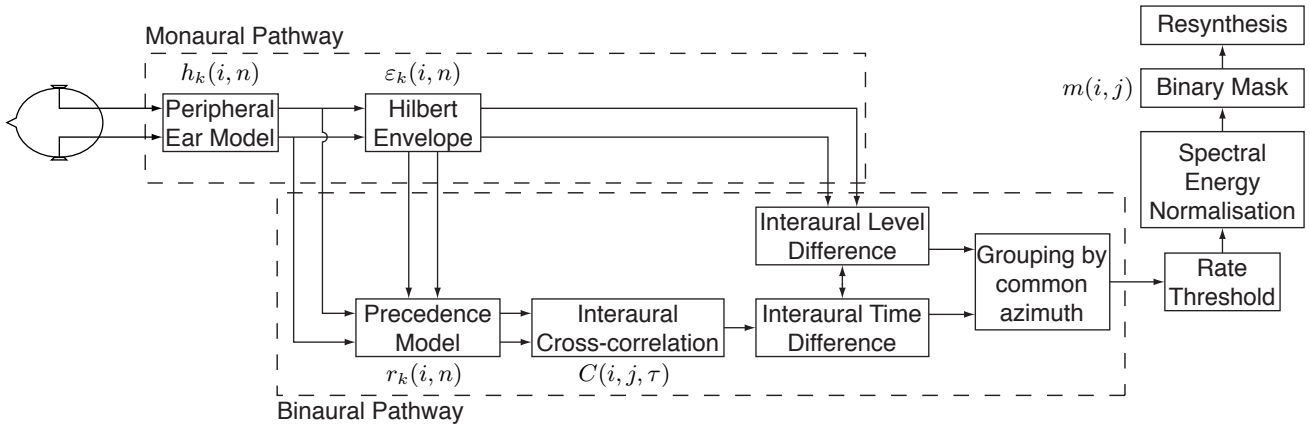

Fig. 1. Schematic of the binaural processor based on [2].

where α_p is the inhibitory time constant chosen to be the number of samples corresponding to 15 ms and A is set to give unity gain at DC.

The target and interferer azimuths are obtained by warping the cross-correlograms to the azimuthal domain, reducing them to skeleton cross-correlograms [2], [14] and summing across time and frequency. The binary mask is calculated from the magnitude of the cross-correlogram at the target and interferer azimuths thus:

$$m(i,j) = \begin{cases} 1 & \text{if } C(i,j,\phi_t) > C(i,j,\phi_n) \\ & \text{and } 10\log_{10}\left(\dfrac{C(i,j,\phi_t)}{\widehat{C}}\right) > \Theta_c \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where ϕ_t is the target azimuth, ϕ_n is the interferer azimuth and

$$\widehat{C} = \max_{i,j,\phi} C(i,j,\phi) \quad (5)$$

Generally Θ_c was set to -160 dB. Following this, two additional checks are performed on the mask by comparing energy values against a rate threshold and by comparing the estimated azimuth against an Interaural Level Difference (ILD) template (see [2] for details).

III. Experimental Procedure

This section documents the procedure used to test the algorithm, including the variables used, choice of metric and signals, and how the Binaural Room Impulse Responses (BRIRs) were obtained.

A. Experimental Variables

The algorithm was tested using a similar procedure to that described in [2]. Specifically, the algorithm was 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
- Three Target-to-Interferer Ratios (TIRs) of 0, 10 and 20 dB (RMS)
- Three interferer signals: white noise, male speech and a modern piece of rock music. Signals are discussed later in this section
- A range of reverberant conditions. The BRIRs were obtained using a different procedure to [2] (see later in this section)

These variables give rise to 135 experimental combinations.

TABLE I
Room acoustical properties.

Room	ITDG [ms]	DRR [dB]	RT_{60} [s]
X	N/A	∞	0.00
A	8.72	6.09	0.32
B	9.66	5.31	0.47
C	11.9	8.82	0.68
D	21.6	6.12	0.89

B. Signals

As stated above, similar signals to those used in [2] were used in the experiment. The target signal was a 4 second excerpt of female speech taken from the European Broadcasting Union Sound Quality Assessment Material [15]. The interfering signals were chosen to be: a rock music track ("Action!" by Razorlight), white noise and an excerpt of male speech also taken from [15]. The speech segments were chosen to incorporate a wide range of phonemes.

C. Binaural Room Impulse Responses

It was decided to use Binaural Room Impulse Responses (BRIRs) captured from real rooms rather than simulating them, due to the generally poor subjective quality of responses calculated using acoustic models. The responses were captured at the University of Surrey from four rooms (later referred to as rooms A–D) of different sizes that exhibit a range of acoustical characteristics. A Cortex (MK.2) Head and Torso Simulator (HATS) and Genelec 8020A loudspeaker were used to capture the responses. The loudspeaker replayed sine sweeps that were deconvolved to produce the impulse responses. For the anechoic condition (later referred to as X), a similar procedure was used and impulse responses were obtained using a pseudo-anechoic approach whereby the responses were captured in a large room and truncated before the first reflection. The Initial Time Delay Gap (ITDG), Direct-to-Reverberant Ratio (DRR) and reverberant decay time (RT_{60}) of each room are given in Table 1.

D. Variation of Precedence Parameters

The algorithm was evaluated in each of the acoustic conditions for a range of values of the inhibitory time constant α_p (see (3)) and the inhibitory gain factor G (see (2)). With $\alpha_p = 0$ or $G = 0$, no inhibition will be triggered and the algorithm will simply cross–correlate the input. The time regions of the input signal that will be inhibited will be affected by varying α_p (i.e. how soon the inhibition

starts after an onset); the strength of inhibition increases with G. Setting these values is a trade-off between selecting reliable regions of the input signal that exhibit minimal corruption by reverberation and maximising the proportion of the input signals that contributes to localisation. Specifically, the input could be highly inhibited with a small value of α_p and a high G; this would yield a signal that is highly uncorrupted by reverberation, but bears little or no resemblance to the input and thus the separation result will be highly inaccurate. Additionally, increasing G will increase the likelihood of cross–correlation values dropping below the grouping threshold Θ_c, resulting in the corresponding T–F unit being excluded at the output. A range of α_p values was used to encompass the range of ITDGs exhibited in the BRIRs: $\alpha_p = [0, 25]$ ms. The range of G values used was based on the value used in [2]: $G = [0, 1]$. The algorithm was first tested by varying α_p with $G = 0.5$ to obtain the optimal time constant for each room. Following this, given the optimal time constants, G was varied to obtain the optimal value for each room. It was expected that the optimal time constant would be correlated with the Initial Time Delay Gap (ITDG) of the room under test, which will vary with the source and receiver positions and with the size of the room, and the optimal gain would be correlated to the Direct-to-Reverberant Ratio (DRR), which will vary with source–receiver distance, with source directivity and with the total acoustic absorption of the room.

E. Choice of Metric

To assess the performance of the algorithm, the widely utilised Signal-to-Noise Ratio metric proposed by Hu and Wang [16] is employed. The version proposed by Hu and Wang uses the target resynthesised from the Ideal Binary Mask (IBM) [17] as the ground truth and is thus termed the Signal-to-Ideal-Noise Ratio:

$$\text{SINR} = 10\log_{10}\left(\frac{\sum_n s^2(n)}{\sum_n \big(\hat{s}(n) - s(n)\big)^2}\right) \tag{6}$$

where s is the target signal resynthesised from the IBM and $\hat{s}$ is the estimated target signal. The IBM is calculated from the clean target and interferer signals thus:

$$m_{\text{ibm}}(i,j) = \begin{cases} 1 & \text{if } 10\log_{10}\left(\dfrac{\delta'_{\text{target}}(i,j)}{\delta'_{\text{noise}}(i,j)}\right) > \Theta_{\text{ibm}} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where δ'_{noise} is the energy of the clean interfering signal, δ'_{target} is the energy of the clean target signal, Θ_{ibm} is a threshold value set to 0 dB and δ' is calculated in the following way:

$$\delta'(i,j) = \big(\delta^{3.333}(i,j)\big)^2, \tag{8}$$

where

$$\delta(i,j) = \Big(\varepsilon'\big(i,(j-1)M+1\big)\Big)^{0.3}, \tag{9}$$

$$\varepsilon'(i,n) = \varepsilon(i,n) - e^{-\tau_\varepsilon}\varepsilon'(i,n-1) \tag{10}$$

and τ_ε is a time constant set in samples to 8 ms. Each result reported later is the mean of the SINRs calculated for the two ear signals.

F. Summary of Experimental Conditions

To summarise, the algorithm is tested in all combinations of the following parameters:

- Three azimuthal separations
- Three Target-to-Interferer Ratios
- Three interferer signals
- Four rooms and a pseudo-anechoic condition
- A range of inhibitory time constant and gain values

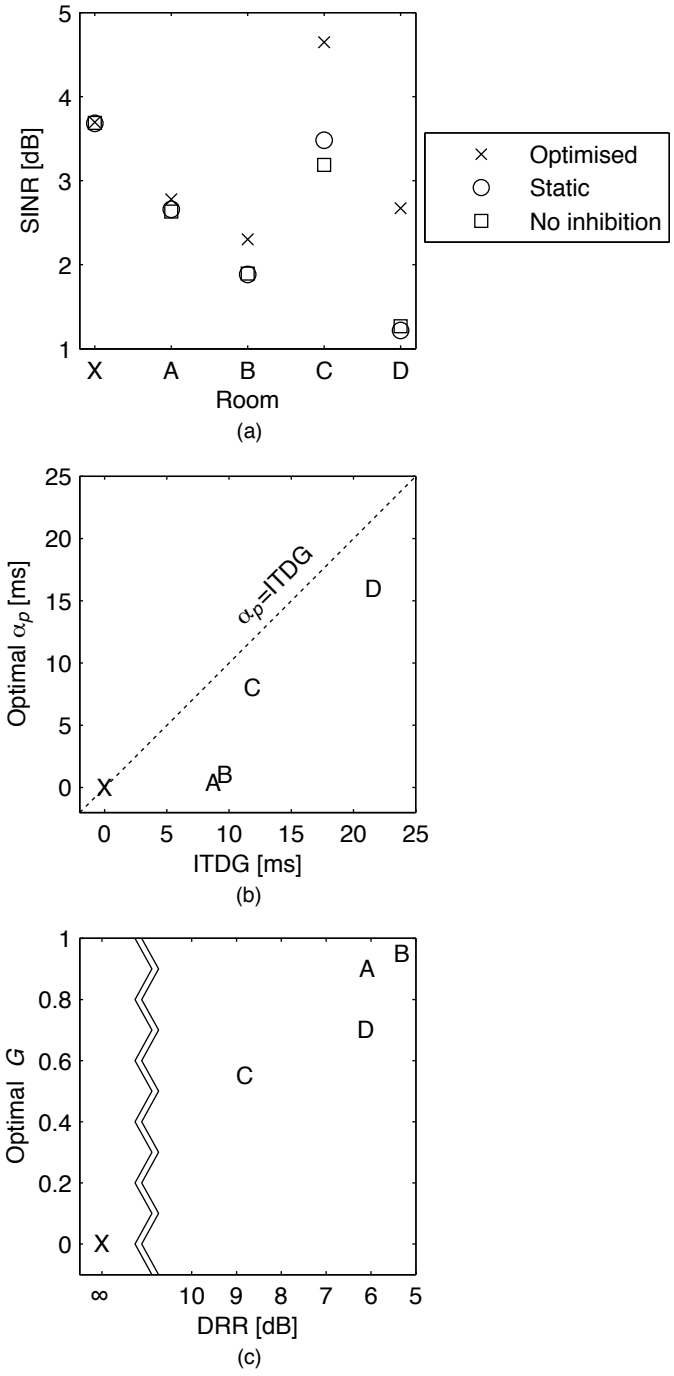

Fig. 2. Modelling the precedence effect. Each result shown is a mean across all of the experimental variables (except the room). (a) The performance of the separation algorithm with the optimised precedence model in each room compared to two other conditions: a static inhibition case ($\alpha_p = 15$ ms, $G = 1$) and with no inhibition ($G = 0$). (b) The value of α_p that achieves optimal performance with $G = 0.5$. (c) The value of G that achieves optimal performance with α_p optimised.

IV. Results and Discussion

The performance of the algorithm across the different rooms is plotted in Fig. 2(a). The results are averaged across all experimental variables, except for the room. The plot compares the optimum performance, obtained by varying α_p and G, with:

- a static condition where $\alpha_p = 15$ ms and $G = 1$ (the values recommended in Palomäki et al.'s original paper [2])

- an un-inhibited condition where $G = 0$

This plot clearly demonstrates the need to adapt the inhibition to the acoustic conditions under test. Furthermore, given that the RT_{60} increases from left to right (see Table 1), the magnitude of performance gain achieved by optimising the precedence model increases with RT_{60}. For the most reverberant conditions, the optimised model produces a significant gain in performance.

In order to implement an adaptive precedence model, it is necessary to identify the acoustic parameters that correlate with the optimal precedence model parameters. It was hypothesised in Section III-D that the optimal inhibitory time constant α_p might be related to the ITDG of the room and that the optimal inhibitory gain G might be related to the DRR of the room. These two hypotheses are tested in Fig. 2(b) and (c) respectively. Fig. 2(b) demonstrates a clear correlation between the ITDG of the room and the optimal inhibitory time constant. Similarly, Fig. 2(c) demonstrates a strong correlation between the DRR and the optimal gain.

These results are in agreement with the aforementioned hypothesis. In terms of the inhibitory time constant, firstly, it is clear that the precedence processing must maximise the proportion of the direct sound that is utilised for localisation. Secondly, it is clear that the inhibition must start before the first reflection, since in all cases the optimal value of α_p is less than its corresponding ITDG. In terms of the inhibitory gain factor, there is a clear compromise between maximising the amount of input signal that contributes to localisation whilst suppressing information corrupted by reverberation. Hence, the optimal gain is related to the DRR, i.e. the relative level of direct and reverberant sound.

V. Conclusions

The Clifton effect is widely observed in psychoacoustics as an adaptive aspect of the precedence effect. Whilst the utilisation of a computational equivalent in source separation algorithms has been suggested, its necessity has not been formally validated.

This paper has shown that, at least for the particular separation algorithm tested, the addition of non-adaptive precedence processing is not necessarily beneficial to the SINR achieved (rooms X, A and B). Furthermore, in certain acoustic conditions (room D), precedence processing can be detrimental to the system performance. In circumstances where the non-adaptive precedence model is of no benefit or is detrimental, an adaptive model can be beneficial (rooms A, D and B). In circumstances where the non-adaptive precedence model is already beneficial, an adaptive model can offer further improvement (room C).

The adaptation of the precedence model is dependent upon at least two acoustic parameters: ITDG and DRR. The ITDG of the room determines the point at which the inhibition should start. The precedence model should maximise the proportion of the input signal's duration that contributes to localisation but ensure that inhibition starts before the first reflection. The DRR of the room determines the appropriate amount of inhibition. The precedence model should suppress information corrupted by reverberation but maximise the proportion of the input signal's amplitude that contributes to localisation. Optimising these two parameters can yield a significant gain in separation performance, especially in more reverberant conditions.

VI. Future Work

This work suggests two areas for future research. Firstly, the necessity for a computational Clifton–like processor in other computational precedence models needs to be investigated. Secondly, a computational Clifton–like processor needs to be developed to determine optimal values for α_p and G automatically based on blind estimation of ITDG and DRR.

VII. Acknowledgements

This work was supported by the EPSRC. The authors would like to thank the anonymous reviewers for their helpful comments.

References

[1] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *The Journal of the Acoustical Society of America*, vol. 106, pp. 1633–1654, Oct. 1999.

[2] 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 Communication*, vol. 43, pp. 361–378, Sep. 2004.

[3] H.-M. Park and R. Stern, "Missing feature speech recognition using dereverberation and echo suppression in reverberant environments," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, vol. 4, 2007, pp. 381–384.

[4] K. Martin, "Echo suppression in a computational model of the precedence effect," in *Applications of Signal Processing to Audio and Acoustics, IEEE ASSP Workshop on*, 1997.

[5] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *The Journal of the Acoustical Society of America*, vol. 116, pp. 3075–3089, Nov. 2004.

[6] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals," *The Journal of the Acoustical Society of America*, vol. 80, pp. 1608–1622, Dec. 1986.

[7] E. A. Macpherson, "A computer model of binaural localization for stereo imaging measurement," *Journal of the Audio Engineering Society*, vol. 39, pp. 604–622, 1991.

[8] J. Braasch, "Localization in the presence of a distracter and reverberation in the frontal horizontal plane. II. Model algorithms," *Acta Acustica*, vol. 88, pp. 942–955, 2002.

[9] R. K. Clifton, R. L. Freyman, R. Y. Litovsky, and D. McCall, "Listeners' expectations about echoes can raise or lower echo threshold," *The Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1525–1533, 1994.

[10] R. L. Freyman, R. K. Clifton, and R. Y. Litovsky, "Dynamic processes in the precedence effect," *The Journal of the Acoustical Society of America*, vol. 90, pp. 874–884, 1991.

[11] C. Hummersone, R. Mason, and T. Brookes, "A comparison of computational precedence models for source separation in reverberant environments," in *128th Audio Engineering Society Convention*, London, May 2010, paper 7981.

[12] P. M. Zurek, "The precedence effect," in *Directional Hearing*, W. A. Yost and G. Gourevitch, Eds. New York: Springer-Verlag, 1987, pp. 85–105.

[13] J. Huang, N. Ohnishi, and N. Sugie, "Sound localization in reverberant environment based on the model of the precedence effect," *Instrumentation and Measurement, IEEE Transactions on*, vol. 46, pp. 842–846, 1997.

[14] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, pp. 2236–2252, Oct. 2003.

[15] (1988) Sound quality assessment material for subjective listening tests. Tech. 3253-E. European Broadcasting Union. [Online]. Available: http://tech.ebu.ch/publications/sqamcd

[16] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *Neural Networks, IEEE Transactions on*, vol. 15, pp. 1135–1150, 2004.

[17] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, P. Divenyi, Ed. Norwell, MA: Kluwer Academic, 2005, pp. 181–197.