PERCEPTION-BASED SIMPLIFICATION FOR BINAURAL ROOM AURALISATION

Hüseyin Hacıhabiboğlu

Centre for Communication Systems Research,
University of Surrey,
Guildford, Surrey, GU2 7XH, UK
h.hacihabiboglu@surrey.ac.uk

Fionn Murtagh

School of Computer Science,
Royal Holloway, University of London,
Egham, Surrey, TW20 0EX, UK
fmurtagh@acm.org

ABSTRACT

Room auralisation refers to the process by which the acoustic response of a room is rendered audible using signal processing techniques. The major part of the computational complexity in a binaural room auralisation system arises from the processing of the early reflections. However, most of the early reflections in a room are suppressed by the auditory system in a variety of psychoacoustical processes such as temporal masking and the precedence effect. This paper presents a perception-based data reduction method based on a mathematical model of the precedence effect. Results of a subjective evaluation in the form of a virtual source-identification experiment are presented. It is shown that it is possible to reduce the total number of reflections by 70% without significantly affecting the localisation acuity.[1][2].

1. INTRODUCTION

The acoustics of a room can be represented by the room impulse response (RIR). A RIR is composed of various elements representing distinct properties of the acoustical properties of the room at a given receiver position for a fixed sound source position. If the sound source is not obstructed, the direct wave front arrives at the receiver position earlier than the reflected waves. The first portion of the reflections (i.e. early reflections) is rather ordered and the temporal density of reflections (i.e. number of reflections per second) and modal density of the frequency response (i.e. number of modes per Hertz) are low. The reflection density and modal density gradually increase and the orderliness decreases for the late reverberation portion. For this reason, early reflections and the late reverberation are generally modelled separately. The former is calculated in geometrical models of room acoustics such as the image-source method (ISM) [2], and the latter is synthetically generated by using a variety of artificial reverberators [3]

Binaural room auralisation refers to making audible the acoutical properties of an enclosure for reproduction over headphones. One of the simplest and most flexible ways of binaural auralisation is by employing parametric digital filters modelling the head-related transfer functions (HRTFs) that define the transfer function from a sound source to the eardrums of a listener. Direct sound and early reflections are processed using HRTF filters to present audible information about their direction. As these filters make up for most of the computational complexity associated with binaural auralisation, a number of early reflections need to be selected in order to implement an interactive 3D auditory display on low-power computational devices such as mobile phones, PDAs etc.

The aim of this study is to provide a perception-based method for selecting a small number of specular early reflections in a geometrical model of room acoutics without significaly affecting the reproduction acuity with respect to the perceived localizational and spatial atributes of the auralised room. The paper is organised as follows. Section 2 briefly summarises the mathematical model of the precedence effect underlying the proposed simplification strategy. Section 3 explains the proposed strategy. Section 4 reports a virtual source-identification experiment that investigates the effect of the proposed strategy on the localisation acuity. Conclusions are drawn and prospects for future work are given in Section 5.

2. A GAUSSIAN MIXTURE MODEL FOR THE PRECEDENCE EFFECT

The precedence effect is the generic name describing a group of auditory functions/phenomena which allow accurate localisation of sound sources in complex acoustical environments such as those found in rooms. In its simplest form, the precedence effect can be demonstrated by the classical experiment carried out in acoutical free-field with two azimuthally separated loudspeakers equidistant from the listener. A broadband click is presented from one of the loudspeakers, and a delayed replica played over the other. Three main effects related to the precedence effect can be observed [4]. (1) *Fusion* refers to the perception of a single fused auditory event at short delays less than 1 ms. (2) *Localisation dominance* refers to the relative dominance of the leading loudspeaker over the perceived location of the fused auditory event when the delay is between 1 ms and 5 ms. (3) *Lag discrimination suppression* refers to the conditions under which the listener is unable to tell whether the lagging sound source is to the left or to the right of the leading sound source. In addition to these effects, the fused auditory event is perceived to be wider.

A mathematical model for the analysis of subjective localisation data under the precedence effect conditions was previously proposed [5]. The localisation responses given to lead/lag pairs were modelled to be sampled from a univariate mixture of two Gaussian distributions. The component means were associated with the perceived directions of each sound source and were related to the original directions and the level of the precedence effect. The standard deviations were modelled to represent the perceptual noise in the auditory path. The model accommodated means of representing and quantifying fusion, localisation dominance, lag discrimination suppression, and the widening of the

[1]The work reported in this paper was supported in part by the EPSRC Portfolio Grant GR/S72320/01,

[2]An extended version of this paper has recently been accepted for publication in the ACM Transactions on Applied Perception [1].

auditory event.

Of particular importance for this paper is the lag discrimination suppression property of the precedence effect. In the model mentioned above, modality of the Gaussian mixture distribution fit to the localisation data represented whether the lagging sound can be discriminated by the listener. This is also an indicator as to whether the lagging source (i.e. the reflection) is a perceptually prominent acoustical cue. It was shown shown that the discriminability of a single reflection is a function of the angular separation between the leading and the lagging sound sources. An exponential model was fit to psychoacoustical data so that a modality function was obtained as:

$$F_{mod}(\Delta_\theta) = 0.5\,|\Delta_\theta|\,e^{-k|\Delta_\theta|-l} \quad \text{for} -\pi/2 \leq \Delta_\theta \leq \pi/2 \tag{1}$$

If $F_{mod} > 2\sigma$, where σ is the response standard deviation for the lead-only condition, the response is bimodal. The response standard deviation (σ) is a measure of localization acuity which represents the variability of the subject response in the single sound source case. As the discrimination of the lagging source is not suppressed, the effect of the lagging sound on the spatial properties of the auditory event is significant. Average values of the constants k and l were found to be $k = 0.0343$ and $l = -0.5722$ for $\Delta_\theta < 0$, and $k = -0.0305$ and $l = -0.6892$ for $\Delta_\theta > 0$ showing a left-right asymmetry in the responses. It should also be noted that the modality function was defined for stimuli with a lead-lag delay of 4 ms.

3. PERCEPTUAL SIMPLIFICATION OF BRIRS

The following selection method is proposed in this paper. The image-source model of the enclosure is calculated for a given listener and source positioning. The image sources are then clustered according to their arrival times with respect to the listener position. Afterwards, azimuth clusters are obtained for each temporal cluster to obtain a basis onto which the selection strategy based on the modality function is applied. Finally, the reduced set of image sources are selected using the modality function defined above.

3.1. Temporal Clustering

The image sources obtained using the ISM model are first clustered according to their relative time of arrival with respect to the direct sound at the listener position. Here, it is assumed that a reflection can take over the role of the direct sound and suppress the localization of subsequent early reflections arriving no later than the temporal threshold, τ_{high}, of precedence effect.

Assume that the listener is positioned at $\mathbf{X}_L = (x_L, y_L, z_L)$, the sound source at $\mathbf{X}_S = (x_S, y_S, z_S)$, and an arbitrary image source at $\mathbf{X}_i = (x_i, y_i, z_i)$. The n^{th} temporal cluster γ_n can be represented as the set of image sources:

$$\gamma_n = \{\mathbf{X}_i : (n-1)\cdot\tau_{high}\cdot c \leq |\mathbf{X}_i - \mathbf{X}_L| - |\mathbf{X}_S - \mathbf{X}_L| < n\cdot\tau_{high}\cdot c,\ n \in \mathbb{Z}^+\} \tag{2}$$

where τ_{high} is the higher temporal threshold (i.e. *echo threshold*) for the precedence effect, and c is the speed of sound. In other words, the early reflections arriving at the listener position ($\Delta t_{n-1} = (n-1)\cdot\tau_{high}$) later than the direct sound but not later than ($\Delta t_n = n\cdot\tau_{high}$) are grouped together. The image sources are thus clustered into concentric spherical shells with the listener at the centre.

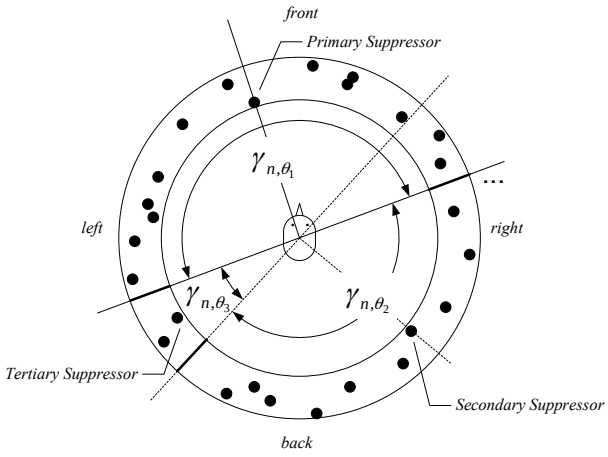

Figure 1: *Azimuth clustering of image sources.*

3.2. Azimuth Clustering

The modality function, F_{mod}, was defined for lead/lag pairs positioned on the listener's horizontal axis for discrimination suppression conditions. If $F_{mod} < 2\sigma_\theta$ for a given pair of image sources, the farther image source (i.e. lagging reflection) is unlikely to contribute significantly to the overall spatial perception if it is not within 1 ms distance of the leading early reflection. Therefore, it is not included in the set of image sources to be auralized. If this condition is not met by the image source, it is included in the set of image sources to be auralized.

A practical problem with using this modality function is that it is only defined for $-\pi/2 < \Delta_\theta < \pi/2$. This necessitates further clustering based on azimuth angle of the image sources within the temporal clusters. This is done in the following way:

In a temporal cluster, γ_n, the image source that is the nearest to the listener position is the primary suppressor of that cluster. Assume that the image source at $\mathbf{X}_\mathbf{n}^{'} = (r_n^{'}, \theta_n^{'}, \phi_n^{'})$ is the primary suppressor of γ_n where radius, r, represents the radial distance from the listener, θ and ϕ represent the azimuth and elevation of the image source with respect to the listener. Using the region where $F_{mod,\theta}$ is defined, the first azimuthal cluster, γ_{n,θ_1}, is formed by grouping together the image sources, $X_j = (r_j, \theta_j, \phi_j)$, for which $\theta_n' - \pi/2 < \theta_j < \theta_n' + \pi/2$. The remaining image sources form the reduced temporal cluster $\gamma_{n^-} = \gamma_n \backslash \gamma_{n,\theta_1}$ ($\backslash$ denotes set difference). The image source in this set which is nearest to the listener position is the *secondary suppressor* positioned at $\mathbf{X}_\mathbf{n}^{''} = (r_n^{''}, \theta_n^{''}, \phi_n^{''})$. The second azimuth cluster, γ_{n,θ_2}, is formed using the same strategy explained above using the reduced temporal cluster, γ_{n^-}. The third azimuth cluster is simply the difference set between the reduced temporal cluster and the second azimuth cluster (i.e. $\gamma_{n,\theta_3} = \gamma_{n^-} \setminus \gamma_{n,\theta_2}$). The image source within this cluster that is the nearest to the listener position is the *tertiary suppressor* positioned at $\mathbf{X}_\mathbf{n}^{'''} = (r_n^{'''}, \theta_n^{'''}, \phi_n^{'''})$. The total number of non-empty azimuth clusters for a given temporal cluster can thus be at least 1 and at most 3 (see Figure 1).

Therefore, the maximum number of non-empty clusters that can be obtained from a set of image sources defining the first T_{res} milliseconds of the room impulse response is $3 \times \frac{T_{res}}{\tau_{high}}$.

3.3. Selection of early reflections

Directional information conveyed in early reflections within 1 ms of direct sound are not suppressed. Such reflections are effective in summing localisation and cause a shift in the perceived auditory event [6]. Therefore, all of the early reflections within 1 ms delay of the suppressor of a given cluster are selected for auralisation. The modality function is used to select the early reflections whose discrimination is not likely to be suppressed by the suppressor of the cluster that they belong to. Consider the cluster $\gamma_{n,\theta_i,\phi_j}$, with a suppressor at $\dot{\mathbf{X}} = (\dot{r}, \dot{\theta}, \dot{\phi})$. For any given image source $\mathbf{X}_i = (r_i, \theta_i, \phi_i)$ in the cluster, which is not within 1 ms temporal distance from the suppressor, the early reflection corresponding to the image source is predicted to contribute significantly to the combined spatial impression of the cluster if the following condition is satisfied:

$$F_{mod} = 0.5 \left| \theta_i - \dot{\theta} \right| e^{-k\left|\theta_i - \dot{\theta}\right| - l} > 2\sigma_\theta \qquad (3)$$

where σ_θ is the standard deviation related to localization blur in azimuth, k_θ and l_θ are exponential model parameters defining the precedence level difference for azimuth and elevation.

4. SUBJECTIVE EVALUATION

A virtual source-identification experiment was carried out to investigate the effects of the perceptual data reduction strategy on the localisation performance.

4.1. Method and stimuli

Image source models for a medium-sized (5 m×7 m×3 m) rectangular room was calculated up to fourth order for seven source positions equidistant at 3 m from a listener positioned at the longitudinal axis of symmetry 2 m away from the back wall. Seven virtual sound sources were situated on the horizontal plane from -15° to 15° with respect to the front direction, with an angular separation of 5° spanning an azimuth angle of 30° in total. All of the absorption coefficients of the surfaces were selected to be equal to $\alpha = 0.3968$ in order to obtain a room with a reverberation time of $T_{60} = 300$ ms.

The binaural room impulse responses (BRIR) were calculated for all the virtual sources using three selection methods. The first set of BRIRs (i.e. *ori*) used all the calculated image sources in the model. The second set of BRIRs (i.e. *beg*) used the image sources selected according to their level as suggested by Begault [7]. Namely, any early reflection with a level 21 dB below the direct sound at a delay of 3 ms and greater, and 30 dB below the direct sound at a delay of 15 ms or greater were eliminated. The third set of BRIRs were calculated using the image sources selected according to the strategy explained in this paper (i.e. *per*). The value for the response standard deviation was selected to be, $\sigma_\theta = 1.5^\circ$ which is a representative value in subjective localisation studies under similar conditions.The numbers of image sources used in the calculation of each BRIR were similar across all of the modeled source positions. The average number of image sources used in the calculation of the BRIRs were 370 (*beg*), 438 (*ori*), and 137 (*per*). The blocked meatus HRTF measurements of KEMAR obtained from the CIPIC HRTF database [8] were used for the calculation of BRIRs.

The stimuli consisted of 500 ms-long frozen white noise bursts. These stimuli were first convolved with the calculated BRIRs. Late reverberation calculated using a feedback delay network (FDN) type artificial reverberator was added to the calculated signals. The stimuli were grouped into blocks according to the model reduction strategy (i.e. *ori*, *beg*, and *per*). The presentation order of the blocks was randomised. The ordering of the stimuli in each block was also randomised. Over the course of each block, each stimulus was repeated 10 times. This resulted in a total of 210 presentations for each subject. The subject's task was to identify which virtual source was active, by clicking the corresponding button on a GUI running under MATLAB. The subjects could only listen to a specific stimulus once. They had to register their responses in order to listen to the next stimulus which was played back after 1 s of the subject's response was registered. No feedback was given to subjects during the test as to whether the source they identified was the correct one.

4.2. Subjects and Procedure

Six subjects (four males and two females; aged 26-32) with normal hearing participated in the experiment. The listening test was carried out in an acoustically treated studio space. The stimuli were played back over circumaural headphones. The presentation level of each stimulus was 65($\pm$1) dBA (SPL) measured near the subject's eardrum. The aim of the experiment was explained to each subject prior to the experiment. A short training on how to use the user interface was given. A test run was then carried out to allow the subject to form an individual source localization strategy. The actual experiment started after the subject was confident with both the experimental paradigm and the user interface. The actual test lasted around 30 minutes for each subject.

4.3. Results

The obtained responses were used in the calculation of two parameters for each subject, $\overline{s}$ and $\overline{C}$ that represent the response variability and the constant localization bias respectively, in a source identification paradigm as suggested by Hartmann *et al.* [9] such that:

$$s^2(k) = A^2 \frac{1}{M_k} \sum_{i=1}^{M_k} \left[R_i - R(k)\right]^2, \qquad (4)$$

and $\overline{s}$ is the rms average of $s(k)$, where $s(k)$ is the localization variability for the k^{th} source, A is the angular separation in degrees between each source, M_k is the total number of trials for the k^{th} source, R_i is the subject's response on the source-index scale in the i^{th} trial, and $R(k)$ is the average response for the k^{th} source. Similarly, the average constant localization bias, $\overline{C}$, is calculated as follows:

$$C(k) = A\left[R(k) - k\right], \qquad (5)$$

and $\overline{C}$ is the average of is the localization bias associated with the k^{th} source, $C(k)$.

Figure 2 shows the responses of each subject in the experiment. It may be observed that the localisation performance is not much different for the different strategies of model selection. Response variability, $\overline{s}$, averaged across all subjects for different selection methods are $\overline{s}_{beg} = 3.57^\circ$, $\overline{s}_{ori} = 3.88^\circ$, and $\overline{s}_{per} = 3.49^\circ$. These values are in general higher than the response variability observed in a real room. The reason for this difference is due to the use of generic HRTFs and lack of head-tracking. The localization biases averaged across all subjects for different selection methods are $\overline{C}_{beg} = -0.32^\circ$, $\overline{C}_{ori} = -0.74^\circ$, and $\overline{C}_{per} = -0.86^\circ$.

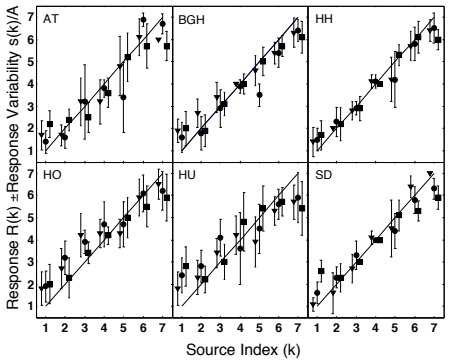

Figure 2: *Localization responses by each subject. Markers: R(k) for* beg (▼)*,* ori (●)*, and* per (■)*.*

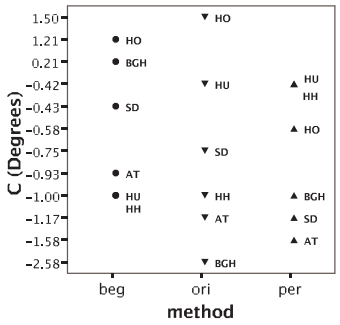

Figure 3: *The constant localization bias, $\overline{C}$ for different subjects and methods.*

Figure 3 shows the localization bias scores for all the subjects. It may be observed that the localization bias is negative for most subjects for all different methods. This suggests a left-right asymmetry biased towards left. Such asymmetry in spatial hearing, particularly with precedence effect experiments, has previously been reported [10]. Two one-way analysis-of-variance (ANOVA) models were fit to $\overline{s}$ and $\overline{C}$ values calculated for the factors *Subject* and *Method*. Post-hoc multiple comparisons were carried out using the Bonferroni correction. *Subject* was a statistically significant factor at the $\alpha = 0.05$ level ($F = 12.74,\ df = 5,\ p < 0.001$) in the ANOVA model for the response variability, $\overline{s}$. Multiple comparisons revealed that the subjects BGH, HH, and SD were better localisers, and that they had lower response variance scores in general. Almost all pairwise comparisons of these subjects with the others were significant at the $\alpha = 0.05$ level. *Method* was not a significant factor. Neither *Method* nor *Subject* were significant factors for the ANOVA model of the constant localization bias scores, $\overline{C}$. This was also verified by the post-hoc multiple comparisons as no statistically significant difference existed between different subjects and image-source selection methods. The results lead to the conclusion that although subjective differences exist, the proposed perceptual simplification method does not have a significant degrading (or improving) effect on the localization performance. The same conclusion also holds for the other selection method (i.e. *beg*). However, the number of selected image-sources is lower for the proposed method which makes it more desirable.

5. CONCLUSIONS AND FUTURE WORK

The work reported in this paper considered the perception-based simplification of binaural room auralisation using the properties of a mathematical model of the precedence effect proposed by the authors earlier. It was shown that it is possible to reduce the number of early reflections to around 30% of the original number of early reflection without significantly degrading the localisation acuity. In another set of experiments investigating the same simplification strategy, similar results were reported for the subjective rating of spatial qualities of the auralised sound field such as presence, spaciousness, and envelopment [1]. The proposed method is based on the selection of early reflections depending on azimuth only. However, it is known that the precedence effect is also observed for different elevations. Among our current plans future work is to quantify lag discrimination suppression for elevated sound sources, and also investigate how the proposed simplification affects the perception of the range of virtual sources. Other than this, we consider combining the proposed method with level-based reduction methods, or a method based on earlier studies of the discriminability of early reflections in rooms.

6. REFERENCES

[1] H Hacihabiboglu and F Murtagh, "Perceptual simplification of binaural room auralization," *ACM Trans. Appl. Perception*, (accepted).

[2] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[3] W G Gardner, "Reverberation algorithms," in *Applications of Digital Signal Processing to Audio and Acoustics*, Mark Kahrs and Karlheinz Brandenburg, Eds., Boston, MA, USA, 1998, pp. 85–131, Kluwer Academic.

[4] 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.

[5] H Hacihabiboglu and F Murtagh, "An observational study of the precedence effect," *Acta Acustica united with Acustica*, vol. 90, no. 3, pp. 440–456, May/June 2006.

[6] J Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, Cambridge, MA, USA, 1997.

[7] D R Begault, B U McClain, and M R Anderson, "Early reflection thresholds for anechoic and reverberant stimuli within a 3-D sound display," in *Proc. 18th Int. Congress on Acoust. (ICA04)*, Kyoto, Japan, 2004, pp. (CD–ROM).

[8] V R Algazi, R O Duda, D M Thompson, and C Avendano, "The CIPIC HRTF database," in *Proc. of 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA*, Oct 2001.

[9] W M Hartmann, B Rakerd, and J B Galaas, "On the source-identification method," *J. Acoust. Soc. Am.*, vol. 104, no. 6, pp. 3546–3557, 1998.

[10] K Saberi, J V Antonio, and A Petrosyan, "A population study of the precedence effect," *Hearing Res.*, vol. 191, pp. 1–13, 2004.