Audio Engineering Society

Convention e-Brief

Presented at the 140th Convention
2016 June 4–7 Paris, France

Binaural auditory feature classification for stereo image evaluation in listening rooms

Gavriil Kamaris[1], Stamatis Karlos[2], Nikos Fazakis[1], Stergios Terpinas[1] and John Mourjopoulos[1]

[1] Audio and Acoustic Technology Group, Dept. Of Electrical and Computer Engineering University of Patras,26500, Greece, gpkamaris@upatras.gr

[2] Dept. of Mathematics, University of Patras, 26500, Greece.

ABSTRACT

Two aspects of stereo imaging accuracy from audio system listening have been investigated: (i) panned phantom image localization accuracy at 5^0 steps and (ii) sweet spot spatial spread from the ideal anechoic reference. The simulated study used loudspeakers of different directivity under ideal anechoic or varying reverberant room conditions and extracted binaural auditory features (ILDs, ITDs and ICs) from the received audio signals. For evaluation, a Decision Tree classifier was used under a sparse data self-training achieving localization accuracy ranging from 92% (for ideal anechoic when training/test data were similar audio category), down to 55% (for high reverberation when training/test data were different music segments).Sweet spot accuracy was defined and evaluated as a spatial spread distribution function.

1. INTRODUCTION

The quality of stereophonic reproduction cannot be accessed via objective metrics since it depends on many variables related to loudspeaker-listening space coupling [1]. This work employs a binaural parameter extraction, analysis and classification approach for typical stereo set-ups in different listening room scenarios and introduces perceptually-relevant evaluation tools for assessing phantom image localization and sweet spot spatial definition. Other relevant qualities such as Apparent Source Width and Listener Envelopment related to spatial impression will not be considered here. Phantom image accuracy is the robustness of perceptual illusion achieved from the summing localization generated by the 2 loudspeaker signals [1] here accessed via the accuracy of the perceived Direction of Arrival Angle (DOA) with respect to the intended image positioning.

Sweet spot accuracy relates to the maximum displacement from the nominal centre position that can result to a predetermined degree of perceived interaural level (ILD) or delay (ITD) degradation [2].

The obvious advantages of introducing a “virtual listener” assessment tool for such scenarios were initially examined in [3]. Today, such an approach becomes even more relevant due to the introduction of complex multichannel reproduction systems [4,5] but also due to the emerging need for objective evaluation of small-size portable stereo devices which often employ room boundary reflections to improve spatial imaging. The work uses binaural modeling to derive interaural cues, the analysis for sweet spot assessment extending the work by Weirstorf and Spors [5] and employs a novel phantom image angle Decision Tree classifier based on a highly efficient semi supervised training stage [6].

2. SIMULATIONS

The tests were based on simulated experiments for a stereo reproduction scenario with 2 loudspeakers at 2m distance and a height of 1.2m facing at 30^0 a virtual listener seated at the ideal sweet spot position (Fig.1). The BRIRs were evaluated via the CATT-acoustic acoustic simulation software [7], utilizing source directivity from CLF files [8].

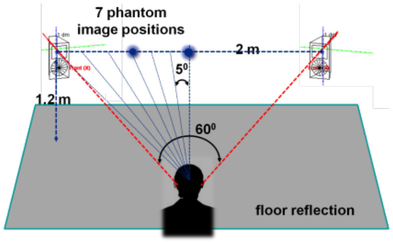

Figure 1 The stereo setup used in the simulations.

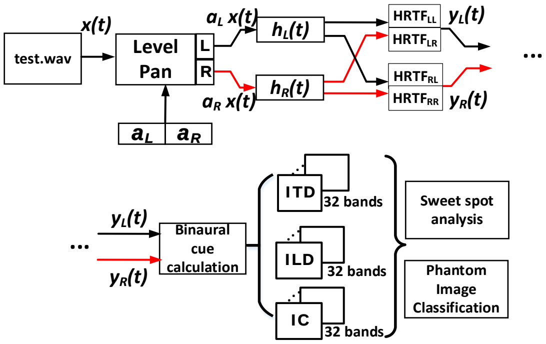

Figure 2 Simulations and binaural cue extraction.

As shown in Fig.2, the binaural signals $y_L(t)$ and $y_R(t)$ were generated according to eq. (1):

$$\begin{aligned} y_L(t) = &\{(a_L x(t) * h_L(t)) * HRTF_{LL} + \\ &(a_R x(t) * h_R(t)) * HRTF_{LR}\} \\ y_R(t) = &\{(a_L x(t) * h_L(t)) * HRTF_{RL} + \\ &(a_R x(t) * h_R(t)) * HRTF_{RR}\} \end{aligned} \quad (1)$$

	Acoustic condition	Source type
T0	anechoic	omni
T1	anechoic	2-way generic l/s
T2	T1 + floor reflection	2-way generic l/s
T3	ITU room (RT=0.25s)	2-way generic l/s

Table 1 Room – loudspeaker test scenarios

The test conditions are given in Table 1. The test signals consisted of white noise bursts with different duty cycle of noise and silence intervals, as well as anechoic audio recordings of speech and music signals. All signals had a total duration of 10 sec. From the signals $y_L(t)$ and $y_R(t)$, a binaural model was employed [9,5] deriving 32 subband ITD, ILD and IC cues per segment (instant).

3. EVALUATION METHOD

3.1. Sweet spot area

Here the sweet spot was evaluated from the DOA angles of the binaural cues (Fig.3) [5]. Thus, at each listening position, the divergence angle (θ_{div}) between the estimate and the ideal one was evaluated, here assumed to be at the nominal central panning position. This divergence angle is compared to the critical angle (here $\theta_{crit} = 5^0$) beyond which a degradation in the image accuracy is assumed to occur [5,9]. The sweet spot area is defined as the listening area for which the divergence angle is not greater than θ_{crit} .

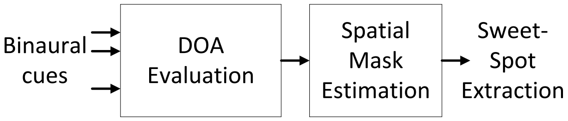

Figure 3 Sweet spot evaluation method

A N-by-N grid is evaluated containing the θ_{div} of each position so that the sweet spot is obtained for the grid points for which $\theta_{div} \leq \theta_{crit} = 5°$. We can define sweet-spot spatial accuracy *A(x,y)*, as the proximity to the ideal angle:

$$A(x,y) = \begin{cases} \frac{\theta_{crit} - |\theta_{div}|}{\theta_{crit}}, & when\ \theta_{div} \leq \theta_{crit} \\ 0\ , & elsewhere \end{cases} \quad (2)$$

Given that every listening position corresponds to a grid area (e.g. *b* in cm^2), we can define the Sweet Spot Area (SSA) in cm^2 as:

$$SSA = b \cdot \sum_x \sum_y A_{bin}(x,y),$$

$$where\ A_{bin}(x,y) = \begin{cases} 1, when\ A(x,y) > 0 \\ 0, elsewhere \end{cases} \quad (3)$$

For the tests in Table 1, the distance between listening points was 10 cm, the grid having 21x21=441 points for a total listening area of 4.41 m^2. For the results in Fig.5, the listening grid starts at 1 m from the loudspeakers.

3.2. Phantom Image classification

Image classification assesses the robustness or deviation of the perceived Direction of Arrival Angle (DOA) with respect to the intended phantom source angle. For the symmetric stereo set-up, the phantom source angles ϑ_{PH}^{i}were generated using the sine panning law [10]:

$$\frac{\sin(\vartheta_{PH}^{i})}{\sin(30^{O})} = \frac{a_L - \alpha_R}{\alpha_L + \alpha_R}\ ,\ 1 < i < 7 \tag{4}$$

Steps of 5^{O} were used creating 7 angle classes from left to centre (Fig. 1). The binaural cues extracted from the signals $y_L(t)$ and $y_R(t)$ were fed to a classifier trained with short segments of the signals. During the evaluation stage, the classifier was driven by cues derived from signals with random and hidden panning and estimated the DOA angle class. The classification accuracy $ACC_N(f)$ gives the ratio of correctly classified instances in the test set to the total number of the test instances *N*:

$$ACC_N(f) = \frac{1}{|N|}\sum_{j=1}^{|N|} I(f(\boldsymbol{X}_j) = \vartheta_{PH}^{j}) \tag{5}$$

where f is the classifier and I is the identity factor, a logical function that gives 1 when $f(\boldsymbol{X}_j) = \vartheta_{PH}^{i}$ and 0 when $f(\boldsymbol{X}_j) \neq \vartheta_{PH}^{i}$. The selected classifier follows the well-known Random Forest ensemble strategy [6]. Apart from the usual supervised procedure where binaural cues from labelled phantom source positions are employed to train the classifier, here an alternative semi-supervised learning method has been implemented [17,6]. Semi-supervised classification (SSC) aims to predict the class ϑ_{PH}^{i} of the unlabeled phantom sources, using only a few labelled examples. Two options for SSC were examined here:

(a) **Self-Training** an iterative scheme during which unlabeled examples along with labelled are exploited, when their certainty exceeds a predefined threshold.

(b) **Co-Training** where the feature vector is split into different binaural cue subsets (e.g. ILDs, ITDs) so that the mutual subset iterative knowledge building functions can develop a more robust and general learning model.

The training / evaluation feature matrix is shown in Fig. 4 and the number of instances *N* was:

$$N = i\ x\ f_s(Hz)\ x\ DutyCycle(s) x\ L \tag{6}$$

where i is the angle classes, fs is the sampling frequency (44.1kHz), *DutyCycle* is the duration of the burst signal cycle (1sec) and L is the labelled data coefficient. For the SSC, L = 0.1 so that cues from only 100ms out of the 10s signals were utilised.

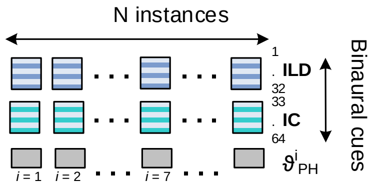

Figure 4 Example of the data structure used for the classifier training and evaluation

4. RESULTS

4.1. Sweet spot area

The results of Fig.5 for noise burst signals, illustrate that the proposed SSA metric is extremely narrow for the ideal anechoic listening case and expands with the addition of early reflections due to room acoustics.

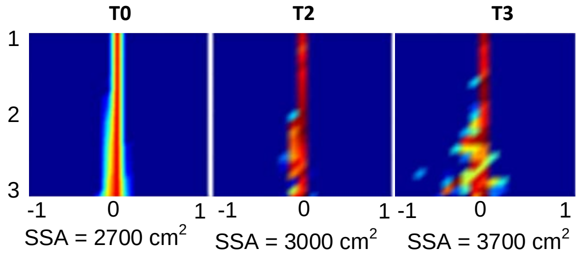

Figure 5 Sweet Spot Area for 3 different test scenarios

4.2. Phantom image localization

For the results in Fig.6, white noise burst training data from the ideal case T0 were used to evaluate phantom image angle of noise burst for listening scenarios T0-T3. Only 100ms out of the 10s signal were employed for SSC training and evaluation.

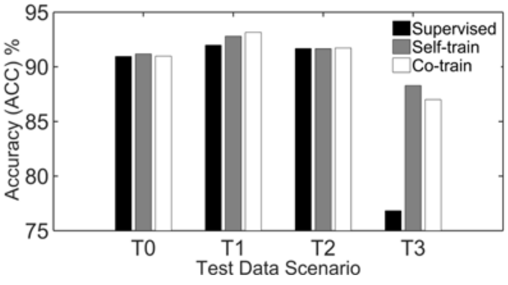

Figure 6 ACC % for different listening scenarios

In Fig.7, evaluation is shown for different audio signals reproduced in the ITU room set-up (T3). Here, the SSC was trained by white noise bursts in the anechoic T0 set-up. When trained by similar signals in the T3 set-up, then ACC results improved for specific signal classes.

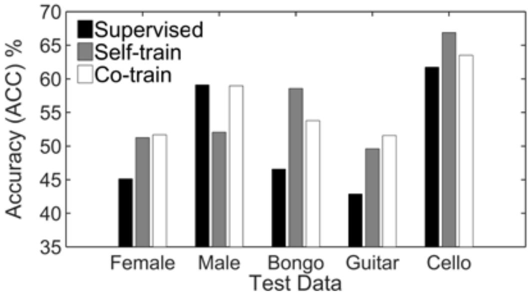

Figure 7 ACC % for different signals for the case T3

In Fig.8 the Chi-squared statistic from all SSC phantom image localization training cases are shown, indicating the significance of each feature on the training process. Note that the frequency-dependent complementarily between ITD and ILD resembles the duplex theory results [15]. Such prior knowledge will be further exploited in future optimization of SSC training.

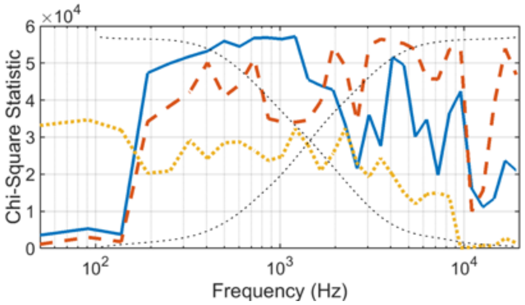

Figure 8 Chi-squared statistic for subband contributions in the SSC training – ITL - - ILD ·· IC and ··· theory.

5. CONLUSIONS

The preliminary study results indicates potential for the proposed objective stereo image evaluation based only on classification of binaural cues without use of higher level perceptual model. The SSC can detect accurately phantom images and as expected, the ACC accuracy reduces with reverberation; in contrast the sweet spot area expands with the addition of early reflections.

Self-train and Co-train SSC options achieve comparable performance, which was superior in most cases to supervised training; significantly, both training and evaluation require binaural features from extremely short signal segments (100 ms) and accuracy is still acceptable when classification is performed for different signals and room conditions to those used for training.

6. REFERENCES

[1] Toole, F. E. Sound reproduction: Loudspeakers and rooms. Taylor & Francis, 2008.

[2] Parodi Y. L, Rubak P., “.Objective Evaluation of the Sweet Spot Size in Spatial Sound Reproduction Using Elevated Loudspeakers, JASA 128(3), 2010.

[3] Theiss B., Hawksford M.J. "Binaural Model-Based Measurements of Phantom Images." AES 105th Convention, 1998.

[4] Harma A., Lokki T., Pulkki V. "Drawing quality maps of the sweet spot and its surroundings in multichannel reproduction and coding." AES 21st Conference., 2002.

[5] Wierstorf H., Spors S., “Predicting localization accuracy for stereophonic downmixes in Wave Field Synthesis,” Forum Acusticum, pp. 1–6, 2014.

[6] Fazakis N., Karlos S, Kotsiantis S., Sgarbas K., “Self-Trained LMT for Semisupervised Learning,” Computational Intelligence and Neuroscience, 2016.

[7] CATT-Acoustic v9.0c – www.catt.se

[8] Common Loudspeaker Format www.clfgroup.org

[9] Dietz M., Ewert S. D., Hohmann V., “Auditory model based direction estimation of concurrent speakers from binaural signals,” Speech Commun., vol. 53, no. 5, pp. 592–605, 2011.

[10] Breebaart, J. and Faller, C. “Spatial Audio Processing: MPEG Surround and Other Applications”, John Wiley & Sons, Ltd, 2007.

[11] Song, Yangqiu, Changshui Zhang, and Shiming Xiang. "Semi-supervised music genre classification" IEEE ICASSP, 2007.