

Towards Individualization of Binaural Music Reproduction

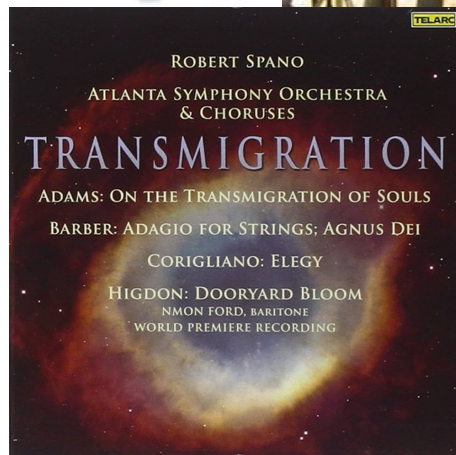
Sungyoung Kim
(with Rai Sato)

Agenda

- Binaural capture of reproduced classical music
- Neural-Network Clustering of Listeners' Hedonic Responses
- Comparison of Binaural Renders
- Discussion
 - Auditory selective attention

Binaural recording

- Dummy head has been used in the music recording more than 50 years.
- AQUA by Edgar Froese (Tangerine Dream), 1974.
- Michael Bishop
 - TRANSMIGRATION (2010)
 - AND MORE ...



Binaural capture of reproduced classical music

- Multichannel-reproduced music provides listeners' with “*a highly precise, natural, and fully immersive listening experience.*” (Hamasaki et al., 2007)
- How to deliver the same experience to people without the multichannel speaker system?



Binaural capture of reproduced classical music

➤ For comparison of various multichannel audio contents

1. Re-record using a binaural microphone (dummy head).
2. Render binaural signals from multichannel contents through convolution of HRIRs (virtual surround).
3. Virtual loudspeaker from HoA signals (captured or simulated Eigenmic, ViReal, Senzi, ...)

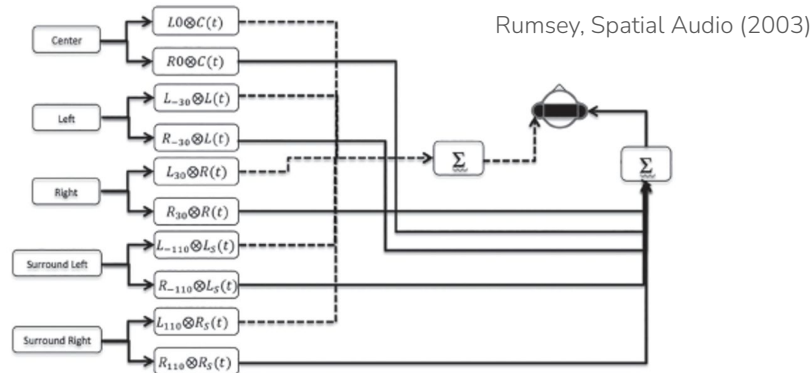
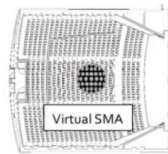
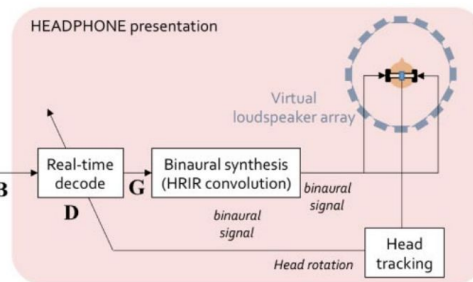


Figure 4.5 Five-channel virtual surround sound implementation diagram.

Sound Field Reproduction



Auralization



Otani et al., 2020, AST

Binaural capture of reproduced classical music

- Dolby ATMOS & Apple Spatial Music
- Will it be another industry-driven yet soon-to-disappear technology mirage (like the 3D-TV boom promoted through the big hit of the movie AVATAR)?
- No one knows the answer but the industry now gives an unprecedented attention to this 'binaural' world.

The Apple Music logo, featuring the Apple logo icon followed by the word "Music" in a sans-serif font.The Dolby ATMOS logo, featuring the Dolby logo icon followed by the word "Dolby" in a bold sans-serif font and "ATMOS" in a smaller font below it.

We focuses on ‘individual difference’
(from socio-cultural backgrounds
and corresponding cognitive styles).



Neural-Network Clustering of Listeners' Hedonic Responses

- Comparison of four listening rooms
 - McGill
 - Tokyo University of the Arts
 - RIT
 - Tohoku Univ. (Anechoic)
- Same B&K 4100 HATS
- 48 listeners
- Two classical orchestra excerpts



1. Recording



4. Binaural reproduction

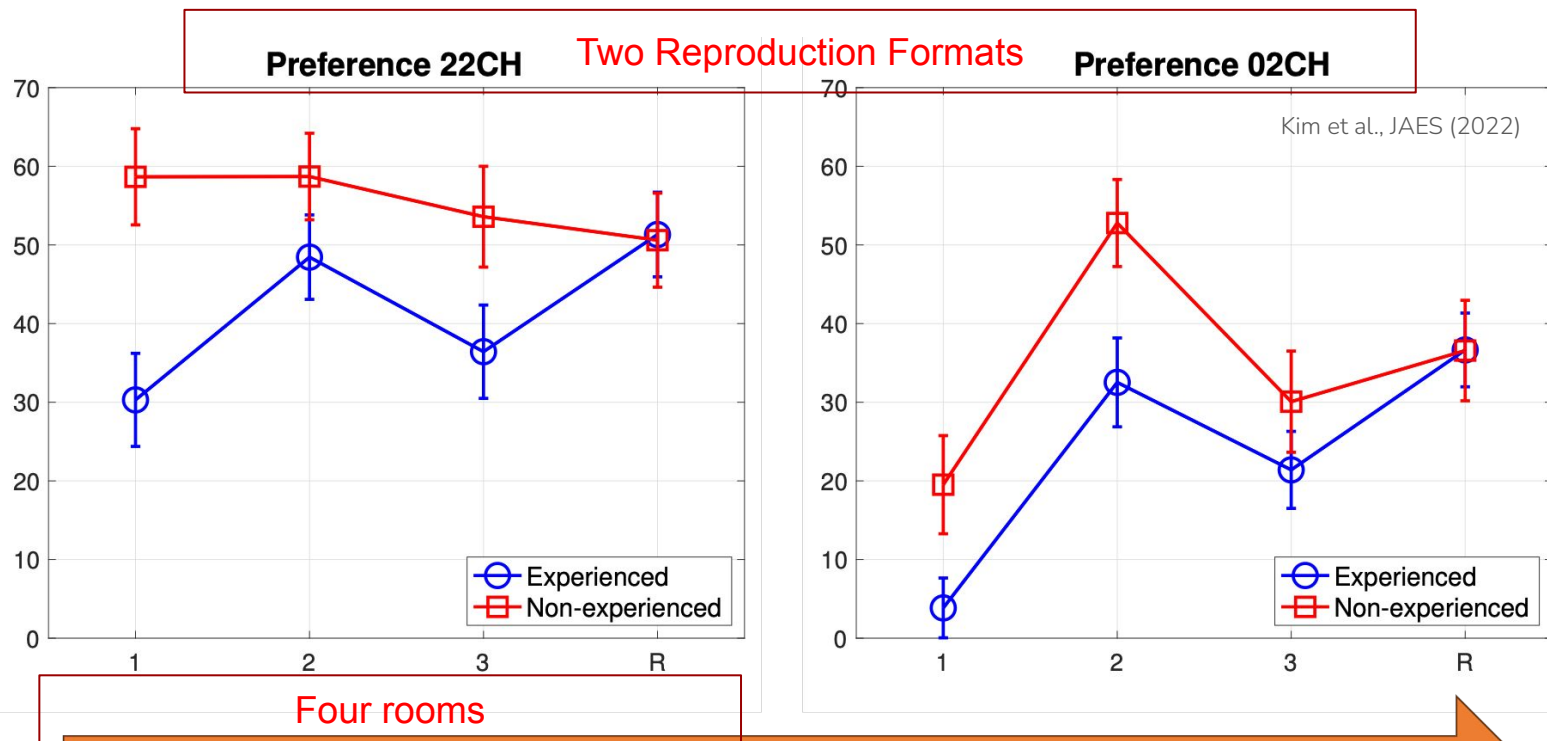


2. Mix for 22.2- and 2-channel



3. Binaural capture

“MANUAL” Clustering of Listeners’ Hedonic Responses



“tailoring systems / services to the needs of specific user groups”??

Neural-Network Clustering of Listeners' Hedonic Responses

- Previously...
 - The inter-subject difference based on their musical experience/training

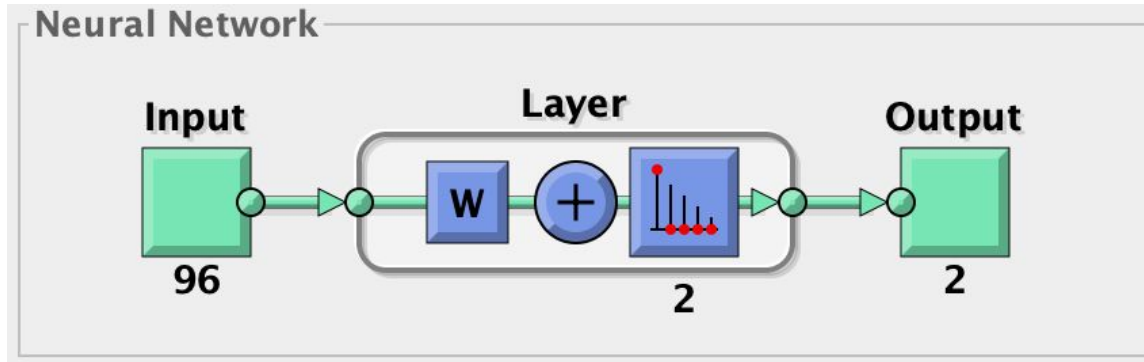
Effect of Skill Level on Listener Performance in 3D Audio Evaluation, JAES, 68(9) p. 628-637 (2020)
Influence of the Listening Environment on Recognition of Immersive Reproduction of Orchestral Music Sound Scenes, JAES, 69(11), p. 834-848 (2021)

- How to judge a subject's musical proficiency?
- Would my 3-year training at Tonmeister school (McGill or Tokyo University of the Arts) validate my proficiency in the evaluation of three-dimensional reproduced music? What about 5-year training?
- A listener may study at an engineering school but could have in-depth understanding in 3D music and immersive auditory experience.
- Data-driven approach, rather than a self-reported or self-evaluated metric?



Neural-Network Clustering of Listeners' Hedonic Responses

- Clustered with 96 features derived from the ratings provided by 101 participants (with new data).
- 96 features: six ratings * four stimuli * four rooms.
- Self-Organizing Map (SOM)
 - a non-linear generalization of the principal component analysis (PCA)



Validation

- Data from 19 participants who were either students or teachers at Tokyo University of the Arts.
- Data from 13 participants from Belmont University.
- The participants were **manually** clustered based on their proficiency in evaluating and manipulating 3D sound fields.
- The cluster process involved analyzing survey results that included questionnaires on their musical training, audio-related education, technical ear training, and other relevant factors.

Confusion Matrix

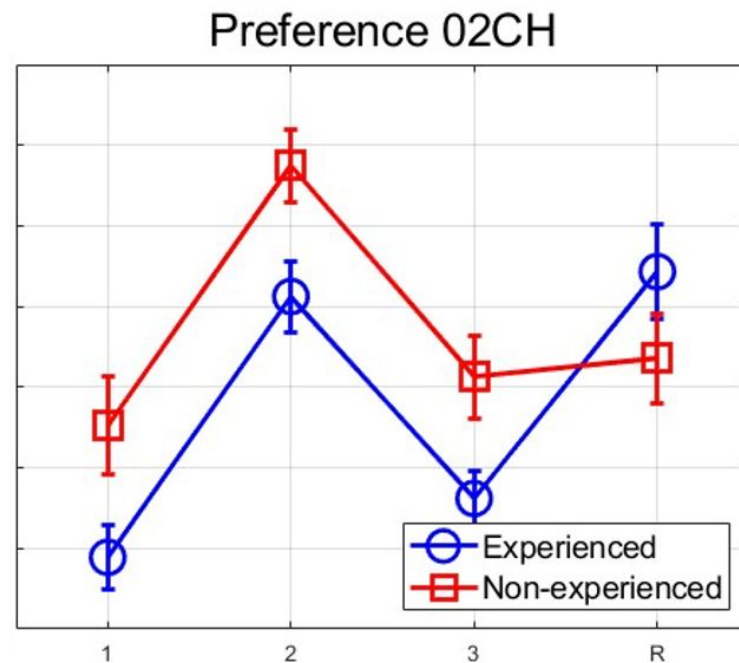
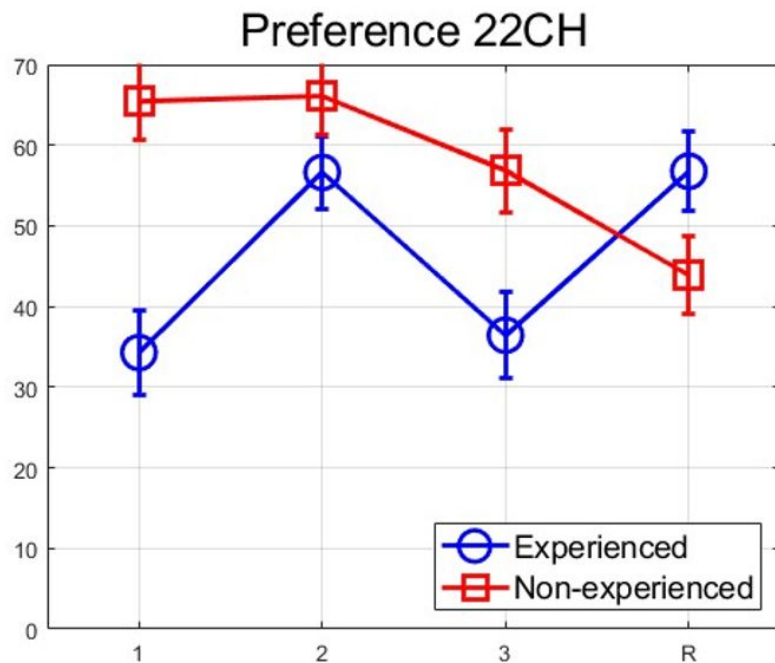
	Positive (actual)	Negative (actual)
Positive (predicted)	True Positive (TP)	False Positive (FP)
Negative (predicted)	False Negative (FN)	True Negative (TN)

Confusion Matrix

	Exp. (actual)	NoE (actual)
Exp. (predicted)	16 True Positive (TP)	2 False Positive (FP)
NoE (predicted)	1 False Negative (FN)	13 True Negative (TN)

- Accuracy $(TP + TN)/(TP + TN + FP + FN) = 29/32 \rightarrow 90.62\%$
- Precision $TP/(TP + FP) = 16/18 \rightarrow 88.9\%$
- Recall (Sensitivity) $= TP / (TP + FN) = 16/17 \rightarrow 94.1\%$
- Specificity $= TN / (TN + FP) = 13/15 \rightarrow 86.7\%$

NN-Clustering Results (126 listeners):



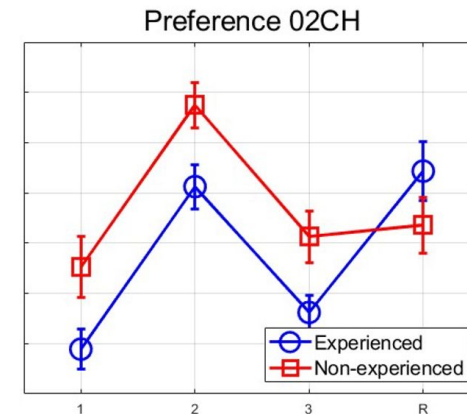
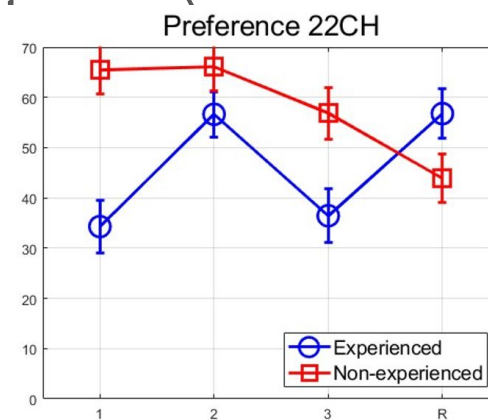
62 Exp. and 64 Non-E.

Can we answer why? → Stepwise Regression

- Group1 (Experienced)
 - **TIMBRE** and **CLARITY** are two significant predictors (R = 0.981) for the mean group preference ratings
 - Timbral aspect more dominant?
- Group2 (Non-Experienced)
 - **ASW** and **DEPTH** are two significant predictors (R = 0.983) for the mean group preference ratings
 - Spatial aspect more dominant?
- Two groups appear to have distinctly different understanding between ASW and LEV.

NN-Clustering Results (126 li

- Room-related reverb was overridden by content-related reverb as for the Group 2 (Non-experienced).
- Contents * BRIR → Enough solution for the Group 2?



ERA of commercial binauralizers

“No need loudspeaker arrays any longer!?”

Binauralizer Comparison

There are many binaural renderers available for immersive audio and VR / AR / Game productions.

However, **these renderers differ significantly in their algorithms and techniques**, resulting in different auditory experiences even for the same piece of audio.

This makes the content creator difficult to select the optimal binaural renderer.



<https://audiophileview.com/headphones/binaural-sound-part-1/>

The purpose of this study

To evaluate the impact of **the systematic differences on listeners' overall preferences and impressions.**

Binauralizer Comparison

Method

28 participants from South Korea, Japan, and US compared 5 binaural renderers and 3 stimuli on 3 attributes.

Attributes

- Overall Preference
- Spatial Fidelity
- Timbral fidelity

Binaural Renderers

- Binaural renderer in Logic Pro
- Virtuoso
- dearVR PRO
- NovoNotes 3DX
- Dummy head (BK 4100D)

22.2ch music clips

- Kaido-Tosei
- Mars
- Lenna

Task: Comparing a pair of two binaural sounds of the same music and rating their perceived magnitudes of the attributes.

The experiment flow was based on Scheffé's Pairwise Comparison (Nakaya's variation).
The collected magnitudes were categorized through the Self-Organizing Map (SOM).

The screenshot shows the experiment application GUI. At the top, it displays 'Subject ID 999' and a 'Submit' button. Below this, it indicates '1 / 34' trials. The main instruction is 'Compare both stimuli and answer how the preference/quality of B is compared to A.' The interface is divided into three sections for different attributes: 'Overall preference', 'Spatial fidelity', and 'Timbral fidelity'. Each section shows two stimuli, 'A' and 'B', with instructions on how to interact with them (e.g., 'A or K key', 'S or L key'). Below each stimulus pair is a 7-point Likert scale with radio buttons. In the 'Overall preference' section, the 'Neutral' option is selected. Below each scale, the user's answer is displayed: 'You Answer is: B is neutral compared to A.' At the bottom, there is a 'Next' button and a '3. Click' instruction. A vertical 'Output' bar is visible on the right side of the GUI.

Experiment application GUI made by Maxmsp

Binauralizer Con

Perceptual Analysis

Group 1 (Experienced):

- distinguish 'preferable' binaural renderer.

Group 2 (Naïve):

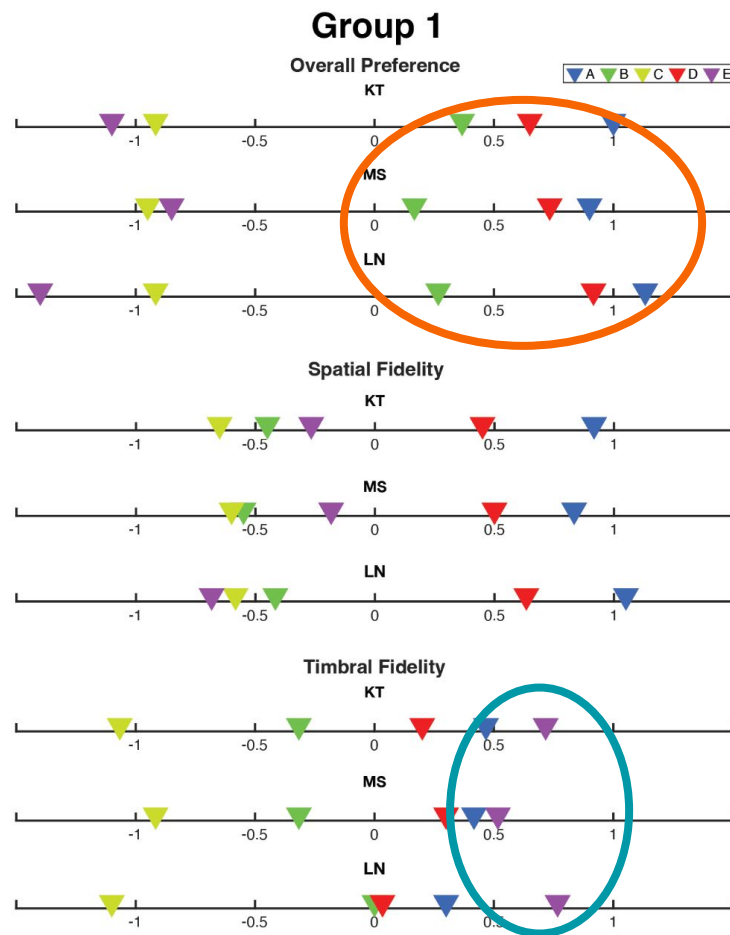
- did not show a clear distinction.

Render A/B/D appeared as preferable:

Integration of the appropriate BRIR may enhance listener preference.

E_(ref.) showed lower preference yet high timbral fidelity:

Complex interaction between spatial and timbral fidelity when assessing binaural renderers.



Conclusion

- A nuanced divergence in listener preferences and sensitivities to binaurally presented music based on their expertise in music and audio production.
- Implications for tailoring binaural rendering approaches for the distinct needs and expectations of different listener groups.
- Refinement and optimization for binaurally presented musical experiences in future.

Thank you for your attention.

Sungyoung Kim

sxkiee@rit.edu

sungyoung.kim@kaist.ac.kr

