The Twenty-fourth Annual Interactive Audio Conference
PROJECT BAR-B-Q 2019

Group Report:
Audio Turing Test

Participants: *A.K.A. "Put the Smoke Back in the Box (aka Enigma)"*
Chris Kyriakakis, USC
David Roach, Magic Leap
Graef Allen, Dolby
Jay DiFuria, Microsoft
Ron Kuper, Bose
David Berol, Amazon
Devon Worrell, Intel
Herv Jasa, On Semi
Nicolas Tsingos, Dolby
Marcus Ryle

Facilitator: Doug Peeler

Problem Statement

Define a framework to evaluate how closely a render system produces an experience equivalent to a live (aka real) one. Additionally, define a methodology to reduce the dependence on human test subjects.

Outline:

1. Scope
    a. As called out in the problem statement, the team's focus was using the framework of an Audio Turing Test to evaluate a render / reproduction system. At first, we focused on the reproduction of a single talker. The target is mostly around single outputs (human voice, triangle, musical instrument, etc.). Machine Learning (ML) should be used to optimize the output system under test.
2. Definitions

Term	**Definition**
ASR	Automatic Speech Recognition
NLU	Natural Language Understanding
ML	Machine Learning
RT60	Time it takes for sound to decay by 60 dB, reverberation time
HATS	Head and Torso Simulator
ERP	Eardrum Reference Point
SNR	Signal to Noise Ratio
dB SPL	decibel, Sound Pressure Level (referenced to 20 µPa)

dB FS	decibel, full scale

3. Assumptions / Caveats

a. Idealized Source Capture (located at source location)
b. Target listener shall also have a listener capture / microphone system (one mic per ear for example)
c. The device under test is not intended to be an Artificial Intelligence (AI) using Natural Language Understanding (NLU) or Speech Synthesis but a reproduction system of a live source
d. Assumption is that test being run can leverage users of the appropriate training
e. The capture system is assumed to be ideal and a practical solution to the capture system is outside the scope of this paper
f. Render system is expected to be targeting a specific use case or product

4. Source Capture System

a. The source capture system shall capture all needed acoustic characteristics of the source. Examples are: frequency response, directivity (which implies multi microphones), level, etc. Additional information such as distance from source will be captured in metadata which will be given to the render system
b. To build an ideal system many aspects some of which are:
    i. Number of microphones and their placement and application (noise floor, directivity, response)
    ii. Source / environment interaction (material properties of walls, distance to walls/corners, RT60, etc.)
    iii. The freedom of motion given to the source and therefore metadata needed to faithfully reproduce
c. Utterance Processing Example for Audio Turing Test v1.0

The below framework is an example of how one could do data collection for an Audio Turing Test. It is not intended as an endorsement of a given brand.

Define the Input of the Recording System

- Four-channel Broadcast Wave Format (BWF)
    - Ch1
        - Name: Primary Source
        - Mic: Neumann TLM103, S/N *TBD*
    - Ch2
        - Primary Reference
        - Mic: Earthworks M30, S/N 9800
    - Ch3
        - Mouth Reference
        - Mic: Earthworks M30 , S/N 9805
    - Ch4
        - Ear Reference
        - Mic: Earthworks M30, S/N 9806
- Each channel:
    - Suggestion: 96 kHz sample rate
        - Requirement: at least 44.1 kHz
    - 24-bit word length
- For each file:
    - iXML data contains details need for cataloging or post processing

Processing steps

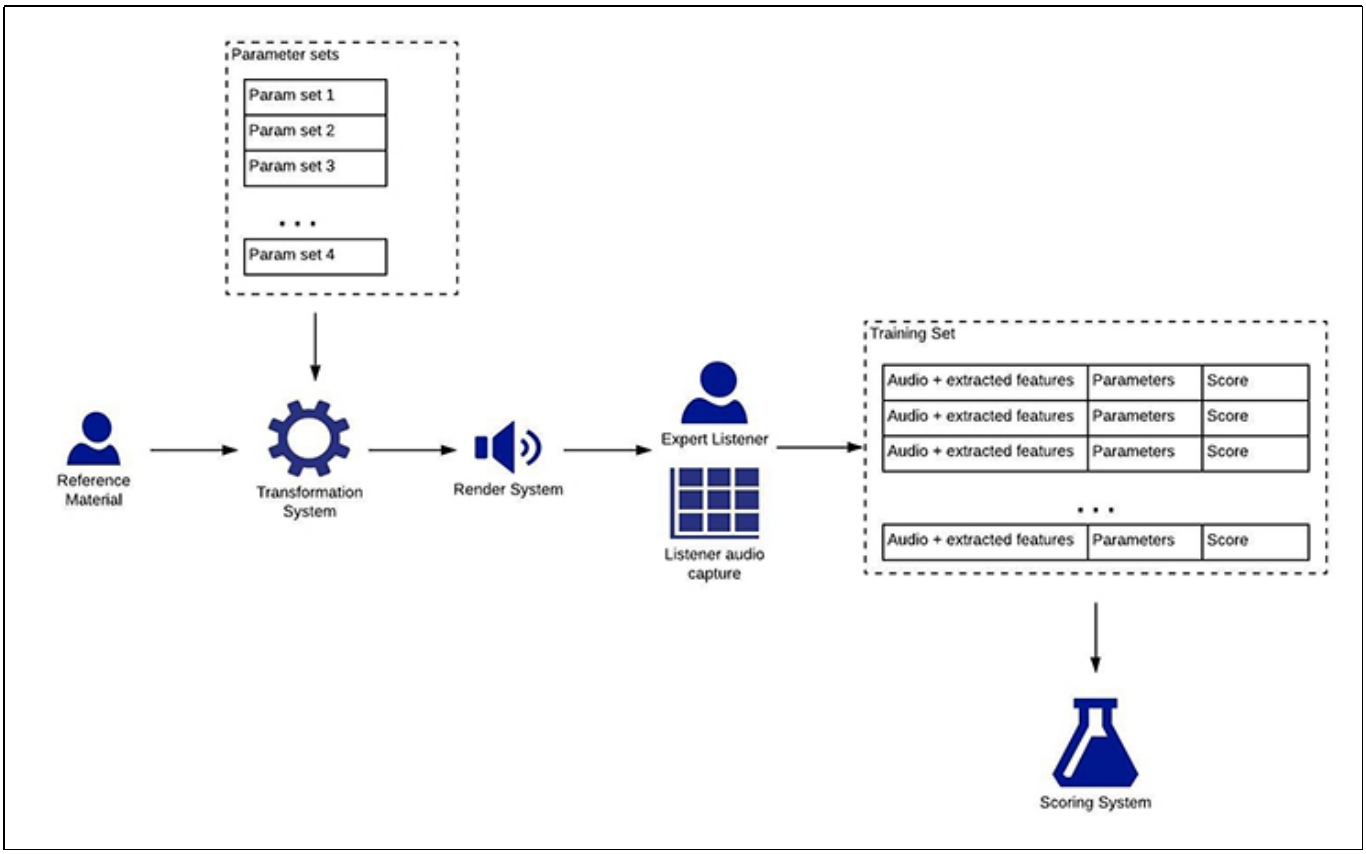
1. Extract each channel from 4-channel file
2. For Primary Source

a. Correct frequency response: gives Equalized Primary Source (EPS)
    i. Use a measured curve captured in situ and in an anechoic chamber
b. Apply high-pass filter to EPS representative to output render
    i. Examples:
        1. Artificial mouths
            a. HMS II.3
            b. B&K 4128
            c. B&K 5128
            d. B&K 4227 or equivalent
c. Check Utterance Level / Crest Factor
d. Normalize levels of HPF-EPS
    i. Currently to –26 dBFS using P.56 ASL
e. Convert to 48 kHz sampling rate, 24-bit word length (for archive)
    i. Suggestion: MATLAB resample at high-quality (e.g. filter length > 128, not default) should be unity gain. If sample rate conversion is not unity gain, do Normalization after SRC
        1. Recommendation: investigate resample software and artifacts that the system may introduce to the playback system
f. Check SNR to validate and reduce reproduction artifacts
g. Measure speech production level
    i. Power analysis on the 96-kHz/24-bit format
    ii. Calibrate/equalize Mouth Reference
        1. use calibration value from iXML for the Mouth Reference mic to calibrate to absolute dB SPL
        2. use the calibration curve for the corresponding S/N for each M30 mic to build an equalization filter for the mic's calibration curve
    iii. For each talker, apply calibration and equalization filter to the Mouth Reference channel, to each file, to get Equalized Mouth Reference (EMR)
    iv. Measure the speech power of EMR using P.56 ASL. This is an estimate of the speech production level, in dB SPL ASL, for that talker on that sentence
    v. Save in a database
h. Measure mouth-to-ear transfer function
    i. Calibrate/Equalize Ear Reference
        1. use calibration value from iXML for the Ear Reference mic to calibrate to absolute dB SPL
        2. use the calibration curve for the corresponding S/N for M30 mic to build an equalization filter for the mic's calibration curve
    ii. For each talker, apply calibration and equalization filter to the Ear Reference channel, to each file, to get Equalized Ear Reference (EER)
    iii. Compute Transfer function using EER and EMR for that talker on that utterance
    iv. Save transfer function in a database
i. Trim 'silence' and avoid background artifacts that would impact reproduction
    i. Fade in and fade out file to avoid discontinuity
j. File naming convention
    i. File name that is human readable
k. Other data for database, from iXML for each file – Optional
    i. Region, Language
    ii. Text of utterance
        1. Compare this text to that produced by ASR in off-line check.
    iii. Talker Gender
    iv. Talker Age Code
    v. Mic S/Ns & Cal values
        1. Cross-check cal values from Cal recordings, also taken for each Day of recording
    vi. Date of recording

5. Transformation System - **the transfer function / system takes the capture system into the**

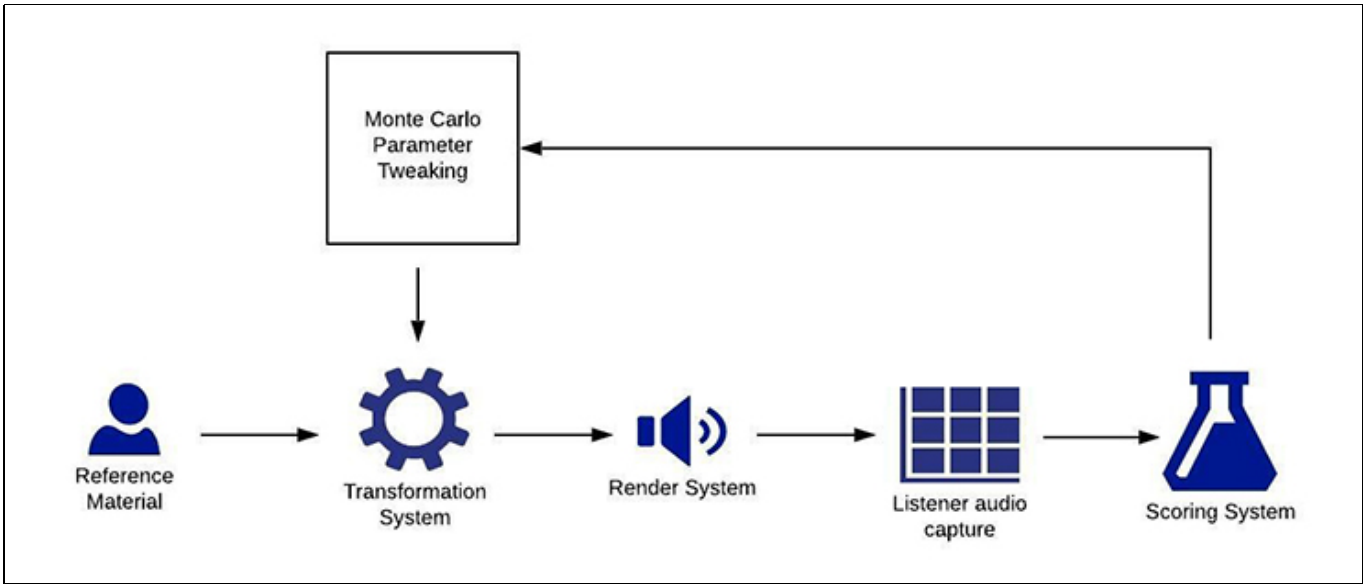
required format for the render system

a. If the Transformation System isn't ideal, it will be part of the Device Under Test in the Audio Turing Test
b. As the input system is ideally defined, we cannot prescribe an exact transformation system but we can articulate the system requirements
c. The owner of the render system is responsible to take the necessary inputs to meet their output render system
    i. If N mics and the system only needs M (M < N) inputs which then maps to P output
d. The Transformation System consumes the capture system metadata and translates and generates new metadata according to the needs of the render system
e. The Transformation System should be part of the Machine Learning Loop to tune the system to reduce the error from the render system



Transformation System Training

This diagram illustrates how ML can be employed to build a scoring system that determines the subjective quality of a transformation system. During the training session, the transformation system is fed a list of sets of control parameters, one at a time. For each parameter set, the expert listener assigns a score such as "0% (terrible)" or "100% (perfect). The listener's score, plus the parameters that were fed into the transformation system, plus extracted audio features from the captured sound, provide the training for an ML based scoring system.



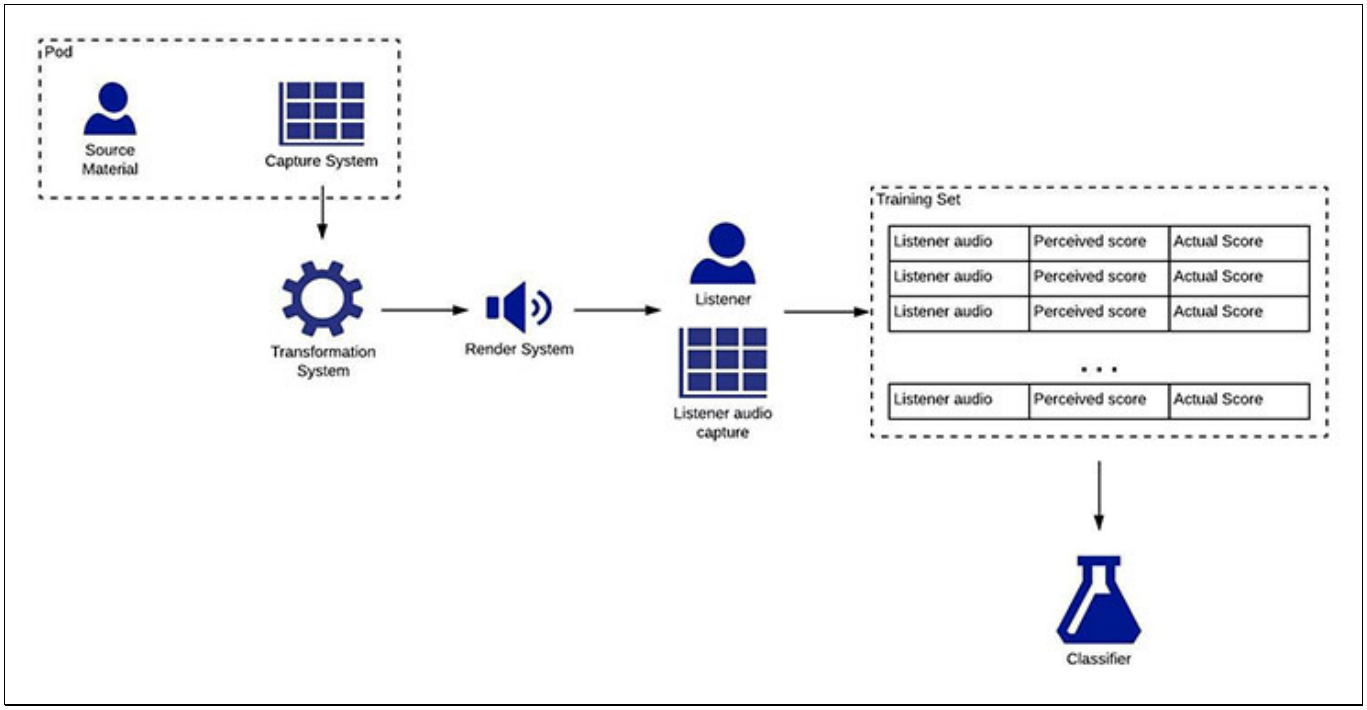
Transformation System Refinement
This diagram illustrates how once an ML-based scoring system has been created, it can be used to continue to refine the transformation system in an unsupervised learning process. Control parameters to the transformation system are varied at random using Monte Carlo methods, and the scoring system is used to accept or reject that random parameter changes.

6. Render System – **is the output system that is under test for believability. Takes the transformation system as an input.**
    a. Different types of render system will have their own requirements
        i. e.g. smart frames or AR will need to have a specific set of setup requirements from that of a speaker system located at the original source location, may have to simulate the room.
7. Listener Capture System – **sufficient number of microphones to capture what is near the user ears (shall be acoustically transparent to not impact the test)**
    a. Render system or Tuning algorithm may require a specific setup
    b. head reference point (acoustic center of head), earDrum Reference Point (ERP), etc.
8. Physical Setup (Person Prep / Training)
    a. Known location could be taken into account or remove location knowledge from source (blindfold)
9. Scoring / Metrics
    a. We are specifically trying to evaluate an Audio Turing Test. Any influence of vision or other senses shall be removed as necessary
    b. [Reference][2] Olive, Sean. (2004). A Multiple Regression Model for Predicting Loudspeaker Preference Using Objective Measurements: Part II - Development of the Model http://www.aes.org/e-lib/browse.cfm?elib=12847
        i. Weighted components for loudspeaker attributes
        ii. Positive correlation
        iii. Abs average deviation 200-400Hz
        iv. Narrow band deviation (100-12kHz octave at a time)
        v. Low frequency extension (–6dB)
        vi. LFQ - Abs deviation from -6dB to 300Hz
10. Human Scoring
    a. Uses all scoring parameters to help automate the process to reduce “cost” of actual test listeners
    b. A-B-X Testing [3] (https://en.wikipedia.org/wiki/ABX_test)
    c. Believability
    d. In the setup of the Audio Turing Test you could re-run against a Mean Opinion Score (MOS) type of score to help tune the system
    e. Additional feedback on the accuracy of the user can be attained through skin resistance and / or EEG sensor.

f. Should presented in a way to minimize “test” pressure

11. Classifier Training

Classifier Training - Listener Evaluation



a. Could use the results of the test to train a neural network to be able to detect what’s realistic (train a classifier to learn from people what’s better or worse.)
b. Could then use the trained ML to figure out how to vary parameters of the transfer function to improve it

12. Metrics / Electrical Characteristics

a. Machine learned automatic improvements

13. Usage and Applications

a. Device types that were considered:
    i. Single Speaker
    ii. Multi-Channel Speaker
    iii. HMD (Headphone, Headset, Earbuds, AR/VR Frames, etc.)
    iv. Consumer Electronics (Phones, Tablets, etc.)
b. Specific Use Cases:
    i. Reproduction System for Voice Assistance:
        1. Prove the render system is a sufficient reproduction of a known human voice.
    ii. Reproduction of Musical Instruments
        1. Single Instrument - Triangle, Violin, Trombone, etc.
        2. Multiple Instruments - String Quartet, Guitar and Voice
        3. Rock Band & Large Orchestra – Can be done but scale and room size may be impractical in most applications
    iii. Naturally Occurring - size and location will need to be considered
        1. Foley type events: Rustling of leaves
        2. Streams and Brooks
        3. Animals, Birds
        4. Wind

14. Clumped Topics / Additional Considerations

a. Deep Fake - How do you know if that sound is real?
    i. Once Synthesis was taken out of scope we didn’t look into this too closely

b. Establishing a set of concrete speaker performance measurements
    i. The render system not passing the Audio Turing Test and the parameters that are tuned to improve performance (either from human or ML system) will be used to understand the concrete speaker performance measurement needed for reproduction
c. Clarity on what makes audio feel Human / Natural
    i. The feedback loop from the ML system would be expected to inform us as to what is required to be perceived as real
d. A human perspective on audio
    i. We are utilizing other senses for validation for the Audio Turing Result but didn't look into what does it mean to be human or organic edge computing

15. Conclusions (necessary?)
    a. Under a bounded setup this is able to be accomplished today. Doing an A-B-X test of a simple acoustic source versus a reproduction system is possible today with this guidance
    b. How to scale to complicated or large sources / rooms will need to be considered
    c. Using ML has a potential to automate the system, both in tuning and in scoring
16. Next Steps
    a. Validation of Audio Turing and the feedback to the reproductive system should be done
    b. A mock-up of the ML that would help with the tuning or scoring systems
    c. Understand the source and room impacts and how to reduce the complexity (and therefore cost) of the system
17. References

[1] Action Item - Utterance Processing Guidelines for Audio Turing Test v0p1 - Dave Berol
[2] Olive, Sean. (2004). A Multiple Regression Model for Predicting Loudspeaker Preference Using Objective Measurements: Part II - Development of the Model. http://www.aes.org/e-lib/browse.cfm?elib=12847
[3] (https://en.wikipedia.org/wiki/ABX_test)