Week 5 Report

What I've done this week

- Read *Learning to Build Natural Audio Production Interfaces (2019)*
    - Article by authors of *A Method for Rapid Personalization of Audio Equalization Parameters (2009)* which I wrote about in last weeks report.
    - Article is an overview of the development of the authors work.
    - The iterative process of learning from the user was refined by using other users feedback as priors.
    - Authors also created a dataset, *SocialFx* (added to datasets doc) which is discussed below.
- *SocialFx* dataset.
    - This dataset is very similar to the SAFE-DB dataset where there are timbre descriptors along with parameter settings (separate datasets for EQ, compression and reverb).
    - Main difference with SAFE-DB is that the listeners performing the tagging were recruited through Mechanical Turk, whereas SAFE-DB users are mostly audio producers.
    - Also parameter settings are for a 40-band EQ rather than a 5-band parametric EQ.
- Read and summarised *Word Embeddings for Automatic Equalization in Audio Mixing (2022)*
    - Short summary and takeaways below.
- Spent comparing the SAFE-DB dataset and plugins with the AudioCommons timbre extractor models.
    - Came across the *Pedalboard* library from Spotify (https://github.com/spotify/pedalboard) which allows VSTs to be loaded directly into Python.
    - Used the above library to apply EQ effects to various audio signals (orchestral music, individual instruments such as guitar, piano and drums).
    - Found that some descriptors have more agreement than others between SAFE-DB and AudioCommons. For example the mean 'bright' EQ setting from SAFE-DB doesn't increase the brightness metric on the AudioCommon model, but the mean 'boomy' EQ setting does increase boominess.
- Watched an interesting talk from the Audio Developer Conference from an *iZotope* engineer around how they apply deep learning to their plugins.
    - https://www.youtube.com/watch?v=fGuT9zoQ_JA
    - Most of their applications seem to be 'offline' to avoid issues with requiring real-time inference.
    - For example, analyzing an input signal for genre and instrument, then applying appropriate EQ settings.
- Read and summarised *Semantic Description of Timbral Transformations in Music Production (2016).*
    - An analysis of the SAFE-DB dataset.
    - Hierarchical clustering is performed on the signal transformation. Shows that clustering correlates with meaningful representations in parameter space.
    - Also shown that some terms are common across different transformation (plugin) types. Though in general, EQ and compression tend to share similar vocabularies for similar transformations, while reverb and distortion have dissimilar description schemas to one another.

Questions

- Would you mind going over what you expect from the Interim Report in terms of breadth of literature review and how defined the research problem should be at the point of submission?
- Also, how much of the implementation of the deep learning model would you expect to be completed by the end of the first semester?

Plan for next week

- Need to consider the inputs/outputs of the model (parameter settings or audio signals) and how this will affect the overall architecture.
- Have a look at whether the SocialFx and SAFE-DB datasets can be combined without compromising the data.
- See if these effects can be applied to general EQ plugins, or whether they will be specific to the plugins used for dataset generation.

Current state of project

- Believe that the SocialFx and SAFE-DB datasets should be enough for the project.
    - Might be good to find if these could be merged into some common parameter language.
- Have a loose high level idea of what the 'workflow' of the model might look like:
    - User inputs a descriptive word/phrase.
    - Model produces some audio effect (one or multiple plugins? Fixed set of plugins or flexible?)
    - User can then make some alteration to these parameters (in the latent space or the actual low-level parameters? Using NLP or controls?)
    - These user preferences are somehow stored for future inference.

Word Embeddings for Automatic Equalization in Audio Mixing

Short Summary

This paper presents a method for generating parameter settings for an audio EQ plugin based on word embeddings of semantic descriptors. These embeddings create meaningful representation of words - even those not present in the training set - to tune a 40-band EQ.

The model architecture itself uses a simple feed-forward network where the input is either a word embedding (obtained from a number of different language models) or a one-hot vector (for the baseline model). The output of the model is 40 nodes which are used to predict the parameter settings for the 40-band EQ - where a value of 0 maps to -4dB and 1 maps to +4dB of gain for that frequency.

The model is trained using cross-validation where the dataset is split into a test and train set and where the test descriptive words are not seen during training. Words are labelled as High Quality (HQ) if they appear frequently in audio mixing literature, and/or as Highly-Rated (HR) if they have a high consistency score for the EQ setting - that is, if the user strongly associated the semantic word with a particular EQ setting. Each test fold contained 9 HQ words and 22 HR words.

To judge model performance, the mean absolute distance was compared between the predicted and actual EQ settings. The authors admitted that this metric was problematic and so the top models were compared using Partial Curve Matching (PCM), which is a method of calculating the similarity between two curves.

A qualitative assessment is also performed by comparing the predicted EQ curves for the baseline, Tok2Vec and GloVe embeddings against the human-tuned EQ curves visually. The embeddings are shown to perform better than the baseline (one-hot) model, but did not always match the human-tuned curve for words in the test set.

Takeaways

- I think that the use of MAE for the loss function is a little bit problematic for a number of reasons.
    - First the error is averaged across EQ settings, then it is averaged across the test fold.
    - MAE treats all EQ bands as equally important, which does not map the human perception of sound.
- The authors mention the SAFE-DB dataset, but do not go to any effort to combine the two datasets (the SAFE-DB dataset uses a 5 parameter EQ). There may be some way to merge these into a common dataset (for example by seeing how the 5-band parametric EQ curve maps onto the 40-band EQ).
- There are a large number of datapoints that are dropped as they have descriptive words in either Italian or Spanish - the authors mention that there are models which may be able to map these onto equivalent English words using a model like ConceptNet.
    - However, there may be further issues with how these semantic meanings translate when describing timbre.
- There is no way of the user refining the output curves without manually adjusting the 40 parameters. This is an area that could be improved (by allowing the user to explore a low-dimensional latent space).