Week 4 Report

What I've done this week

- Gathered a number of papers related to quantifying timbral descriptions of sound and had a quick skim of a few of them.
    - From what I have read so far, there doesn't seem to be a real consensus on how semantic descriptions of sound map to audio features.
        - For example, in one paper [1] UK English speakers were shown to disagree with English speakers in the US on the acoustic properties of the words "warm" and "clear".
    - One difficulty is sonic descriptors can have different meanings depending on context:
        - A "bright" bass guitar may have a EQ boost at 1kHz.
        - Whereas a higher-pitched instrument may require a boost at a much higher frequency band for the acoustic quality to be described as "bright".
- Watched YouTube presentation of a model that uses a VAE to simplify EQ controls called *flowEQ*.
    - Video here: https://www.youtube.com/watch?v=qbMYTLCP4Cg
    - Code here: https://github.com/csteinmetz1/flowEQ
    - Maps 13 parameters down to 1-3 controls in the latent space.
    - Uses SAFE-DB dataset for training (added to dataset doc and discussed below).
- Did some analysis of the *SAFE-DB* dataset:
    - This dataset is the result of the implementation of 4 plugins (reverb, EQ, compression and distortion) where users can upload their parameter settings along with a description of the sonic qualities of the setting (e.g. "warm", "tinny" etc).
    - The dataset is quite small (1700 EQ, 468 compressor, 441 reverb and 309 distortion settings). However, it looks like it has been used in quite a few ML papers around modelling timbral attributes.
- Clarified points about *Black-Box DDSP* paper:
    - Spent a few hours going through the code and implementing the custom gradient function on some toy examples (can go through worked example at meeting).
    - f: this takes in an audio signal x and transforms it to a predicted audio signal $y=f(x;\hat{\theta})$ where $\hat{\theta}$ are the predicted parameters from the encoder layer.
        - Loss is a weighted sum of L1 distance in time domain and frequency domain between target and predicted output (with some accounting for phase shifts).
    - *Input*: for the encoder, the input is a log-scaled Mel-spectrogram. For the Fx-layer it is the raw audio signal and the predicted parameter values from the encoder.
        - These are taken with a frame size of 1024 and 256 sample hop size.
        - Mel-spectrogram also given larger context window (40960 samples)
- Did some more reading into the *AudioCommons* timbre extraction models. The implementation details of the model are slightly vague as there is no output paper, however there is an evaluation report:
    - The models use data from the FreeSound website (https://freesound.org/) which is a large collection of royalty-free audio clips.
    - Experts hand select a high and low anchor example for each of the timbre descriptors.

    - Participants in a listening test are asked to rate sounds subjectively based on the current descriptor.
    - Features are extracted from the audio signals and a regression model is used to determine the feature combinations to use to predict the subjective ratings.
    - Objective measures are used between predicted and mean rating per sample to judge model performance (Pearson/Spearman correlation etc.).
- Read and summerised *A Method for Rapid Personalization of Audio Equalization Parameters (2009)* (takeaways below).
- Read (no summary):
    - *SAFE: A System for the Extraction and Retrieval of Semantic Audio Descriptors (2014)*
        - Paper for SAFE-DB dataset.
    - *Timbral Attributes for Sound Effect Library Searching (2017)*
        - *AudioCommons* paper on collecting a dictionary of terms for timbre descriptors - used for the Timbre Explorer model training.

Questions

- Can use the AudioCommons timbre model as there is no peer-reviewed research paper that has been published? Or is an evaluation report not enough to justify it's use?

Plan for next week

- Read some more papers on quantifying timbral descriptions.
- Look to see how the SAFE-DB dataset has been used in subsequent papers and if there are any issues that have been noted or if data augmentation has been performed.
- Start *Generating Sound with Neural Networks* tutorial series.

Current state of project

- Still searching for suitable datasets - particularly if timbre descriptions need to be used.
- Learning how audio pipelines are implemented in neural networks.

A Method for Rapid Personalization of Audio Equalization Parameters

Short Summary

This paper proposes a method of simplifying parametric equaliser interfaces by mapping high-level user preferences for sound manipulation to parameter settings. The user first inputs an audio signal to be transformed and a descriptive term. A number of probe equalization curves are then generated by concatenating 2-8 Guassian functions, which are used to alter the gain of 40 frequency bands. These are sampled to maximise the difference in EQ curves across channels and presented to the user, who then rates how well the generated EQ curve matches the descriptive term given using a slider between -1 (opposite) and 1 (perfectly matches the descriptive term).

The user evaluations are used to compute a weighted function that represents the influence of each frequency channel in capturing the essence of the descriptive word. This weighted function then gives a new EQ curve, which the user can scale with a slider to amplify the effect. It was found through a listening

test (19 participants) that the weighted function reached asymptotic performance after around 20-30 user responses (<2 minutes).

Takeaways

- I think it is interesting that this method takes into account each individual's subjective opinion of the acoustic qualities of descriptive words.
- There doesn't seem to be much practical use for the descriptive word other than reminding the user what sonic quality they are looking for.
    - It doesn't seem as though these preferences are automatically stored for future use - though the user could just save the final EQ curve as a preset.
    - The method also doesn't learn long-term user patterns for descriptive words - something that might be possible with an ML approach.
- Overall I'm not sure how much long-term practical use there would be for the method.
    - 2 minutes per EQ curve and descriptive word could add up quickly if the user is using this method for a whole recording session worth of instruments.
    - User may learn more by spending the time manually adjusting the EQ parameters.
    - However, might give beginners a good starting point that they can work from.

References

[1] Disley, Alastair C., and David M. Howard. "Spectral correlates of timbral semantics relating to the pipe organ." *Speech, Music and Hearing 46* (2004): 25-39.