Machine Learning for Classifying Pedagogical Methods in Classrooms

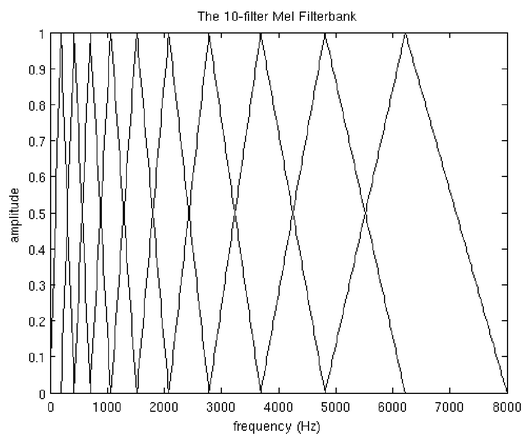
PennState
Kayla Montgomery, Vince Trost, Drew Wham
Teaching and Learning with Technology
The Pennsylvania State University

Abstract

There is current ongoing research into active learning pedagogies and the classic lecture format. Some research has suggested that active learning is better for STEM courses. In this poster, we outline the beginning steps to evaluate the type of pedagogical method being utilized in current courses at Penn State by analyzing audio records. As a proof of concept, we demonstrate that our approach effectively distinguished recordings of lectures from nature sounds, city noises and music. Our approach begins with deconstructing the audio into Mel-Frequency Cepstral Coefficients and other numeric variables. We then employ a tree based classification machine learning algorithm (XGBoost). We then evaluated on held out audio of individual lectures. Our results support the conclusion that this approach was effective at distinguishing between the four audio source classes. We are now gathering and labeling data from classrooms to apply this approach to its ultimate use case.

Features Collected

We extracted multiple features from audio data including Mel-Frequency Cepstral Coefficients (MFCC). MFCC are derived characteristics of audio signals that are widely used in speech recognition.



When an acoustic signal is to be used for speech recognition, the acoustic power spectrum is first transformed into the nonlinear Mel scale, which simulates human hearing perception. A cepstrum is the inverse Fourier tranform of the logarithm of a signal's power spectrum. A Mel-frequency cepstrum is a cepstrum produced from a mel-transformed acoustic power spectrum. The Mel-Frequency Cepstral Coefficients are the amplitudes of the discrete cosine transform of all mel frequency log powers.

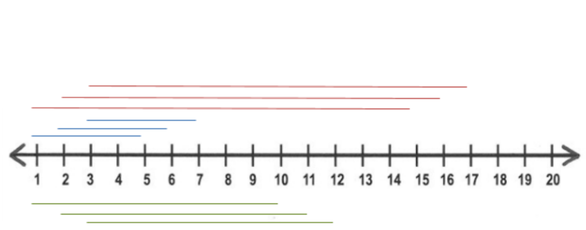
Feature ID	Feature Name	Description
1	Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame.
2	Energy	The sum of squares of the signal values, normalized by the respective frame length.
3	Entropy of Energy	The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4	Spectral Centroid	The center of gravity of the spectrum.
5	Spectral Spread	The second central moment of the spectrum.
6	Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
7	Spectral Flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8	Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9-21	MFCCs	Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
22-33	Chroma Vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
34	Chroma Deviation	The standard deviation of the 12 chroma coefficients.

List of the 34 features extracted using the Python script

Methods

32 hours of audio data were scraped from YouTube. 8 hours each of nature, music, city and lecture audio were obtained. The music audio was obtained as 3 large compilations. The lecture data was obtained from 24 discrete videos of single-speaker lectures, each of which was reduced to a 20 minute excerpt. The nature and city datasets were each derived from individual 8-hour-long ambient audio compilations.

The data was then labeled and run through the Python script that gives the features. The Python script yielded 136 total variables: 34 features each for summarizations of audio data at 1, 5, 10 and 15-second intervals. Different types of audio might be best resolved at different window sizes. Audio interval wasn't quite as important for the lecture dataset used here, but may be more important when evaluating two-speaker data.



The data that was summarized over time would not start at the first second. For example, the columns that had summarizations for the 5 second mark, the first four rows would just be padded with zeroes.

While withholding one lecture at a time from the training data, the XGBoost script was run to give predictions on each lecture excerpt using the withheld video as the testing data.

Results

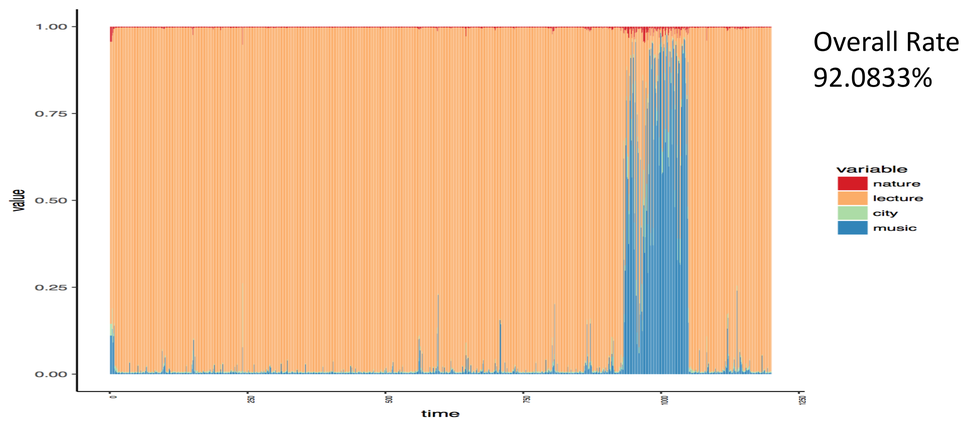
A prediction is performed at each interval. The graphs to the right in the "Audio Classification" section illustrate this concept: a prediction occurs at each second, with a cumulative prediction for the entire audio clip represented by the "Overall Rate".

Most of the lecture excerpts were successfully predicted to be lectures by the XGBoost algorithm. Some of the video lectures had substantial portions classified as nature or music. This is probably due to the presence of unusual sounds not encountered in other lecture recordings. The three music samples were successfully predicted to be music when individually withheld. Both nature and city samples were successfully predicted. Since the source data for these samples were single compilation videos, we could only withhold subsets of the data rather than discrete portions, so this was at best a sanity check.

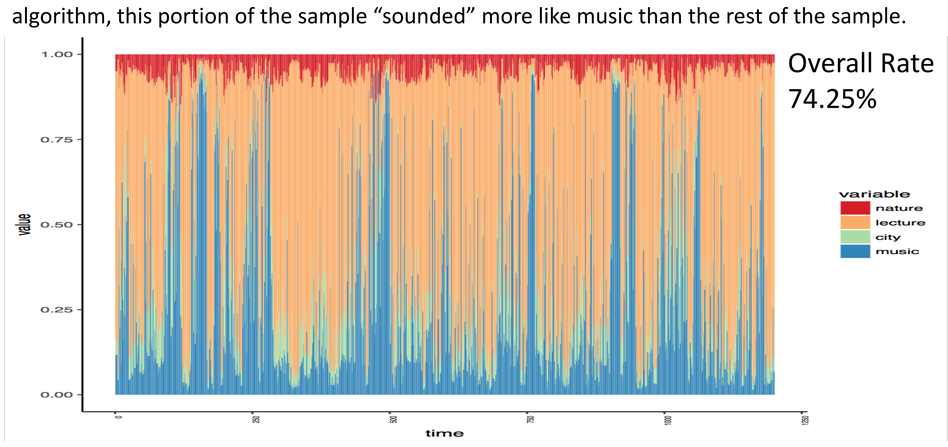
In the excerpts shown in the "Audio Classification" section examples 2 and 3 would likely have had much better predictions if a smoothing function has been applied. The overall correct classification rate for all lecture excerpts was 94.8%. 14 of the 24 lecture excerpts had a 100% correct overall classification rate; no individual interval in these excerpts was predicted to be in one of the non-lecture audio classes.

Audio Classification for 3 Lecture Excerpts

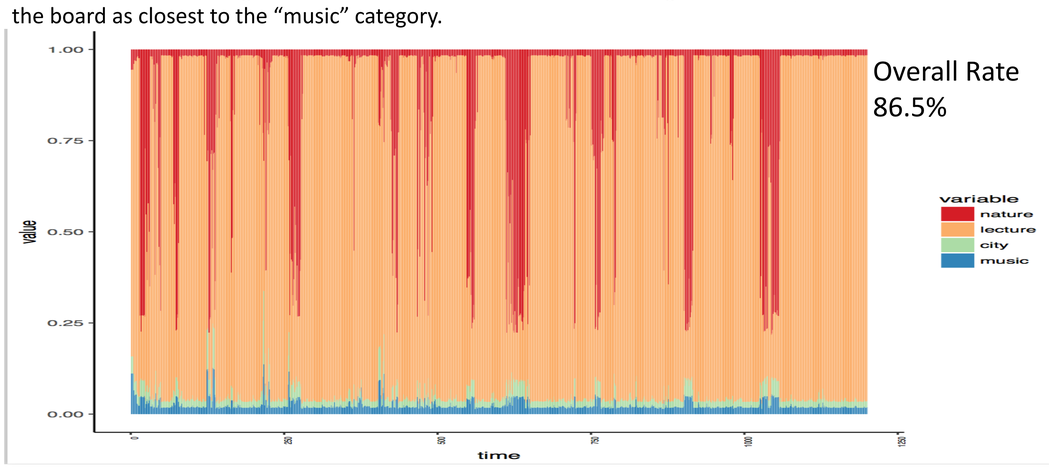
For each second of these sample lecture recordings, XGBoost assigns a percent probability of the audio belonging to each of the four audio classes (nature, lecture, city or music).



Excerpt 1: For this lecture excerpt, most of the audio was classified as "lecture" except at the 850 second mark. Manual review of the lecture excerpt found that at this point in time the lecturer turned on a video. To the algorithm, this portion of the sample "sounded" more like music than the rest of the sample.



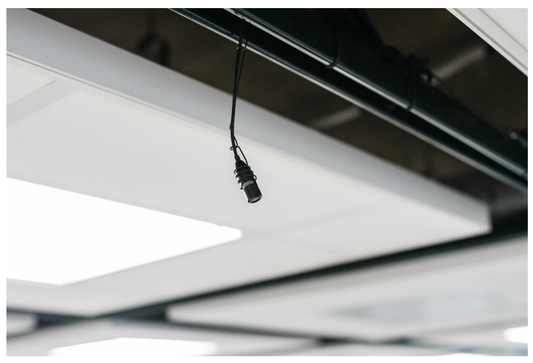
Excerpt 2: This excerpt was less consistently classified. Upon review, it was observed that the lecturer spent much of the time writing on the chalkboard. The trained XGBoost algorithm interpreted the sound of chalk on the board as closest to the "music" category.

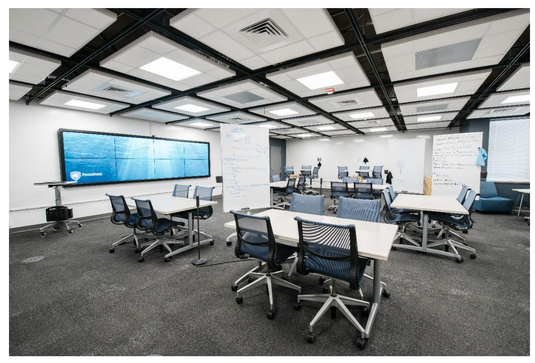


Excerpt 3: It was less obvious why the algorithm classified some periods as most likely to be "nature" sounds. At one point, the lecturer reads a poem with markedly different speech patterns; she later tells a joke which the audience laughs at. Both of these periods were categorized as most likely to be in the "nature" category.

Future Research

We are currently trying to classify types of audio within lectures. Potential categories include single speaker (professor lecturing), two speaker (question-and-response between lecturer and student), silence (students using iClickers), and many speakers (group work). Privacy concerns can render it difficult to permanently record lectures, so machine learning may be valuable for classifying audio data in real time without retaining it.





An array of five microphones in a classroom are used to record audio data.

Contact info: Kayla Montgomery | knm5291@psu.edu | Teaching and Learning Teaching and Learning Technology Department of The Pennsylvania State University
Acknowledgements: The writers of "Classroom sound can be used to classify teaching practices in college science courses" and the Python Audio Analysis library contributors Theodoros Giannakopoulos