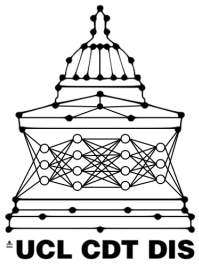

UCL CDT DIS Note

24th May 2021

Audio-Enhanced Segment Retrieval Within the Spotify Podcasts Dataset

Abigail Alexander[a], Matthijs Mars[a], Josh Tingey[a], and Haoyue Yu[a]

[a]University College London

The Spotify Podcasts Dataset, a corpus of 100,000 podcasts, was released in 2020 to promote research into the emerging field of spoken audio, which lags behind other fields of study, such as purely text-based Natural Language Processing. To encourage the use of the raw audio within this domain, we have extracted a series of high-level features from the complete dataset and made them readily available for others to use. Furthermore, using these features, we constructed three audio-based metrics to find podcast segments that were funny, consisted of a debate, or contained disapproval. With these metrics, we have implemented an audio-enhanced podcast segment retrieval system that improves upon a text-based baseline for queries requiring both topical relevancy and an appropriate mood, out of entertaining, subjective, and discussion.

Contents

1 Introduction **3**
1.1 The Dataset . . . 3
1.2 The Tasks . . . 3
1.3 Our Task . . . 4

2 Feature Extraction **5**
2.1 The Geneva Minimalistic Acoustic Parameter Set . . . 6
2.2 YAMNet . . . 8
2.3 Extracted Features . . . 10

3 Metric Engineering **11**
3.1 Creating Labelled Data . . . 11
3.2 Creating Mood Metrics . . . 12
3.3 Performance and Limitations . . . 14

4 Segment Retrieval **15**
4.1 Baseline retrieval . . . 15
4.2 Audio-enhanced retrieval . . . 16
4.3 Retrieval results and discussion . . . 17

5 Conclusion **18**

6 Future Work **18**

Bibliography **19**

1 Introduction

Podcasts are spoken word digital audio files that have been rapidly growing in popularity over the last twenty years. In comparison to other audio media such as broadcast news, podcasts possess a diverse range of content and styles and thus offer a unique opportunity to study and explore speech and language [1]. This is significant because, unlike the well established and somewhat saturated field of Natural Language Processing (NLP), the spoken audio field is an emerging and underdeveloped area of research. This report presents our attempts to probe this research domain by investigating a large dataset of podcast audio files supplied by the audio streaming provider Spotify.

1.1 The Dataset

In 2020, Spotify released a large dataset of podcasts[1] with the view of promoting research into the field of spoken audio (particularly podcast speech) [7]. This dataset contains 105,360 English language podcast episodes collected from Spotify between late 2019 and early 2020. Each episode includes the raw audio (sampled at 44.1 kHz), a transcript generated using Google's Speech-to-Text API, and associated metadata, which includes the episode name, description, publisher, as well as the RSS feed link. In total, the dataset contains approximately 60,000 hours of audio and 600 million transcript words, corresponding to a total of 2 TB worth of data. The large size and lack of associated labels make the datasets practical usage challenging.

1.2 The Tasks

Alongside the release of the dataset, Spotify introduced a new podcasts track to the 2020 Text REtrieval Conference (TREC)[2]. TREC is an ongoing series of workshops that concentrate on various information retrieval research areas, named tracks. The tracks can be divided into several distinct groups: question answering, subject-specific search, and conventional web search. TRECs mission is to foster and promote research in the above fields by providing the necessary infrastructure for large-scale evaluation of various methodologies. The conference collects and presents a standardised dataset, retrieval conditions, question sets, and assessment methods, while participants are expected to develop the associated retrieval system and submit their results within a prescribed time frame.

The TREC 2020 podcasts track was split into two tasks: segment retrieval and summarisation [8]. For segment retrieval, the task consisted of retrieving the most relevant two-minute-long segment from the complete podcasts dataset given a query that could come in one of three types: finding a segment about a certain topic, re-finding a segment you have heard before, and finding information about a known item. The summarisation task consisted of generating a relevant and informative short textual summary of a given podcast.

In this work, we focus on the topical form of the segment retrieval task. This particular form of query is focused on the broad exploration of the complete dataset, with no particular podcast or episode segment in mind. For 2020 each topical query consisted of a title (a short phrase or set

[1] `https://podcastsdataset.byspotify.com/`
[2] `https://trec.nist.gov/`

of words) and a more detailed (paragraph long) description. The expected retrieved segments were required to be two minutes long starting on minute boundaries (e.g. 0.0-119.9 seconds and 60-179.9 seconds) meaning each segment overlaps its neighbours by one minute. This type of segmentation leads to a total of 3.4 million possible segments within the dataset.

Last year none of the segment retrieval task participants used the audio data for retrieval (apart from improving the transcription). Instead, all used the Speech-to-Text generated transcripts exclusively. This choice is understandable as topical relevancy is expected to be heavily driven by the spoken words within each segment. This year (2021), however, the segment retrieval task will be updated to make the information contained within each segment's audio more essential to the task. It is also hoped that this change makes the retrieved segments more engaging, opinionated, and conversational.

Given similar topical titles and descriptions as last year, an additional *mood* category will be requested for each query. The task then becomes the retrieval of the most appropriate podcast segment for an arbitrary combination of topic (title and description) and mood. The three moods chosen are: entertaining, subjective, and discussion, each of which is detailed in Table 1. For reference, some complete example queries are shown in Table 2. Although the text features will remain important for determining the topical content of each segment, it is hoped that the audio should provide additional information on the mood of the spoken audio.

Mood	Description
Entertaining	The topic is presented in a way which the speakers intend to be amusing and entertaining to the listener, rather than informative or evaluative.
Subjective	The speaker or speakers explicitly and clearly express a polar opinion about the query topic, so that the approval or disapproval of the speaker is evident in the segment.
Discussion	The segment includes more than one speaker participating with non-trivial topical contribution (e.g. mere grunts, expressions of agreement, or discourse management cues ("go on", "right", "well, I don't know ..." etc) are not sufficient).

Table 1: The three types of query mood explored within this project. For each topical query, the retrieved segment is also expected to be one of the three moods.

1.3 Our Task

Our task and the focus of this report can be divided into two main aims. Firstly, we wish to encourage more researchers to utilise the Spotify podcasts audio data when participating in the TREC podcasts track by making the audio data more accessible for potential candidates. This can be achieved by supplying pre-computed and understandable high-level audio features in a simple and lower storage size format such that users do not have to manage the large and unwieldy raw audio files that may be otherwise off-putting. Secondly, we want to improve upon a baseline implementation for the TREC segment retrieval task ourselves by using these computed audio features to devise metrics that can help categorise the mood of podcast segments.

Topic Title	Topic Description	Mood
Near death experiences	I wonder if people have shared near-death experiences in podcast episodes. I would like to find and listen to some stories. I am not interested in the science of near-death experiences.	Discussion
Black lives matter	What do people mean when they say "black lives matter"? I am interested in personal reflections that give context to the phrase "black lives matter" and why it is important to individuals. News stores about Black Lives Matter protests are relevant as well.	Subjective
Halloween stories and chat	I love Halloween and I want to hear stories and conversations about things people have done to celebrate it. I am not looking for information about the history of Halloween or generalities about how it is celebrated, I want specific stories from individuals.	Entertaining

Table 2: Examples of different topical segment retrieval query possible at the 2021 TREC segment retrieval task, with title, description, and mood shown.

2 Feature Extraction

As all the podcasts within the dataset are sampled at 44.1 kHz, a 30-minute long podcast contains approximately 80 million values of raw data, a large number from which to extract useful information. This is illustrated in Figure 1 which shows an example of the raw audio waveform from a podcast with both speech and music segments. While there is some difference seen between the different segments, this is mainly due to the loudness of the section with music. Therefore, to extract useful information from the podcast audio, the raw data needs to be processed into more informative high-level features that act to summarise the raw audio. Here, we discuss two different approaches to extracting such features.

The first approach focuses on features that come from established foundations in phonetics and speech sciences. These features include measurements in the time domain (e.g. speech rate), the frequency domain (e.g. pitch), the amplitude domain (e.g. loudness) and the spectral energy domain (e.g. relative energy in different frequency bands).

The second approach uses learned features that are inferred from a labelled dataset. This method typically requires a pre-processing step employing a machine learning model to map the raw audio data to labelled categories. An example of such a dataset is AudioSet [5], which contains audio from approximately 2 million 10-second YouTube clips, labelled by humans into 632 audio event classes. The complete AudioSet ontology covers sounds such as humans, animals, music, amongst many other common everyday environmental sounds.

To provide a large set of useful high-level features which are both understandable and cover a broad range of use cases, we have extracted two complementary feature sets. The first set uses the first approach of established features in phonetics and speech sciences. The selected features

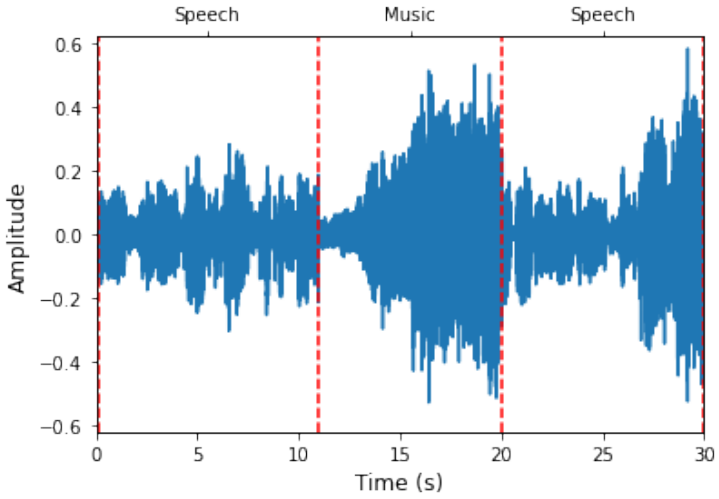

Figure 1: The raw audio waveform of a 30 second clip from a podcast containing both section of both speech and music.

are part of the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [3]. The second set of features uses the second approach of learned features, employing the Yet Another MobileNet (YAMNet)[3] architecture to label the audio data with the labels from the AudioSet ontology.

Since the YAMNet network was designed to label audio in segments of 0.96 seconds in duration, we generated the GeMAPS features using a similar format. The segments were calculated at every 0.48 seconds such that all the segments overlap both their neighbours by half their length.

2.1 The Geneva Minimalistic Acoustic Parameter Set

The Geneva Minimalistic Acoustic Parameter Set or GeMAPS is an attempt by Eyben et al. [3] to design a minimalistic and standardised parameter set for acoustic parameters that are useful for machine learning problems.

A minimalistic parameter set is one that calculates the least amount of features required to generate strong results. This way, classifiers trained on this feature set will not over adapt to the training data but instead generalise well to unseen data. Additionally, the interpretation of the parameters is much easier within a minimalistic set because the features and derived models are relatively simple.

Furthermore, while much research bases its methods on the established features from phonetics and speech sciences, the exact form of calculating these features varies. Therefore, GeMAPS means to provide a set of parameters that are calculated in a clearly defined, standardised manner, such that the results are reproducible and results can be easily compared.

The recommended parameter set contains two parts; one minimal parameter set and an extension to this set. The features consist of both Low-Level Descriptors (LLDs) and features that are

[3] The YAMNet model can be found on `https://github.com/tensorflow/models/tree/master/research/audioset/yamnet`

derived from these LLDs. The so-called functionals are derived from the LLDs by aggregating them over an extended period of time.

The minimal set contains 18 LLDs consisting of features that describe vocal features like intonation, stress, rhythm, excitation, as well as various spectral descriptors that analyse the base frequency and harmonics of speech. The selected descriptors are chosen based on their importance from previous research results. This minimal set is referred to as GeMAPS.

The extension to the minimal set contains seven additional LLDs, all cepstral descriptors, which are features for analysing the periodic structures in frequency data [2]. These features have been shown to consistently improve results on automatic affect recognition tasks with respect to the features in the minimal set of GeMAPS. The extension of this set in combination with the minimal set is called the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS).

In designing this recommended parameter set, the GeMAPS authors compared both the minimal and the extended set with large-scale brute-forced baseline acoustic feature sets on binary arousal and binary valence classification. The results show that eGeMAPS always matches or outperforms GeMAPS, which indicates that the added features can help in predictive tasks. Classification with eGeMAPS achieves similar results to the large scale parameter sets but is overall slightly behind in performance, yet the size of the parameter set is only 2% of the largest set included in the comparison [3].

The eGeMAPS features are available through the openSMILE feature extraction toolkit [4]. The openSMILE toolkit is a tool for Speech and Music Interpretation by Large-space Extraction (SMILE) and contains feature extraction algorithms for speech processing and music information retrieval. The openSMILE toolkit is also available for python[4]. Figure 2 shows a subset of the eGeMAPS features for the same podcast segment as shown before in Figure 1.

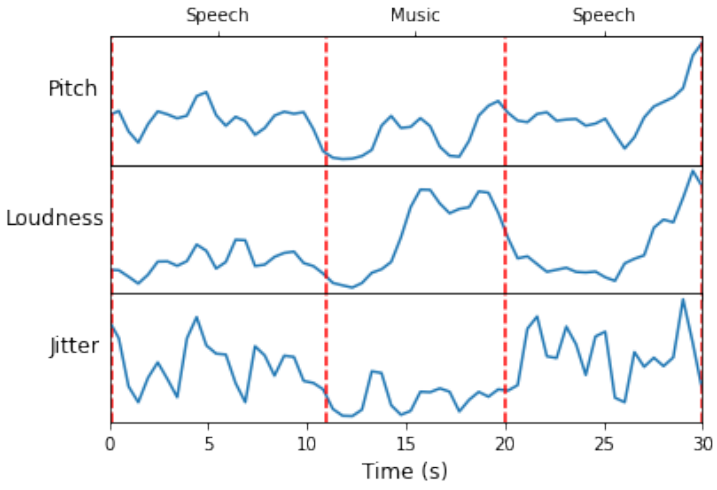

Figure 2: The pitch (F0semitoneFrom27.5Hz_sma3nz_amean), loudness (loudness_sma3_amean) and jitter (jitterLocal_sma3nz_amean) averaged over 0.96 seconds of audio for a clip of 30s from the podcast dataset.

[4] `https://github.com/audeering/opensmile-python`

We decided to extract the eGeMAPS features because of their minimal size and excellent relative performance. Moreover, the standardised way they are generated makes them easily comparable to other research in the field.

2.2 YAMNet

Yet Another MobileNet or YAMNet is a neural network based audio feature extractor based on the Mobilenet_v1 convolutional architecture [6], which is trained on 521 audio event classes from the AudioSet corpus.

The network processes raw audio input into mel spectrograms and uses convolutional neural networks (CNNs) to process these spectrograms in a similar manner to image recognition tasks. A mel spectrogram shows the power in the mel spectrum bins as a function of time. The mel spectrum is a transformation of the frequency information of a audio signal, which is representative of how humans perceive the pitch of sounds. The logarithmic mel spectrogram is a method for visualising periodic structures in the frequency domain. Figure 3 shows the mel spectrogram for a podcast segment from the dataset.

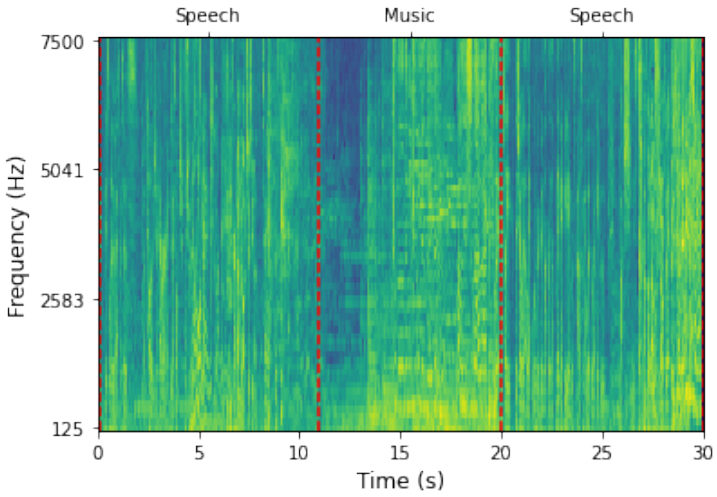

Figure 3: The logarithm of the mel spectrogram for a 30 second clip from a podcast containing both speech and music.

The output of the MobileNet architecture is averaged into a 1024-dimension embedding vector. Then a single logistic layer is used to predict the 521 AudioSet event labels from the embedding vector.

Besides using the scores of the YAMNet as an audio event classifier, the 1024-dimension embedding vector can be used as a feature vector for other purposes. The YAMNet is a feature extractor in this way, and a network can be trained using the embedding vectors and a small set of labelled data, to predict audio events without having to train a very large network. Figure 4 shows the embedding vectors for a podcast segment from the dataset.

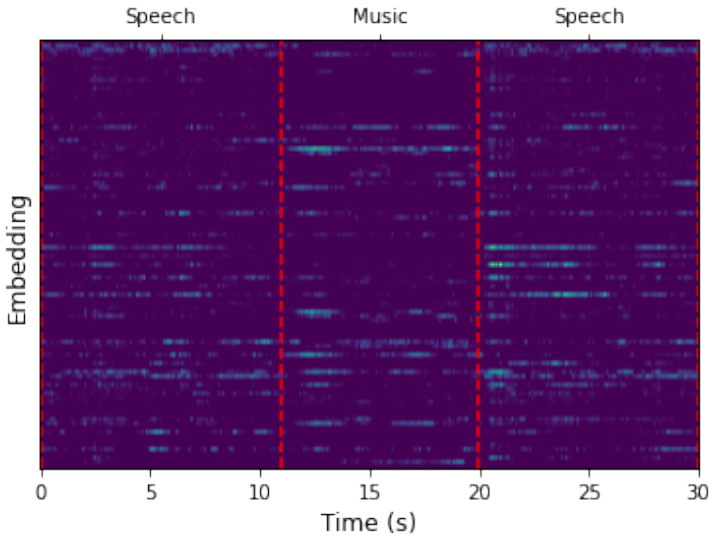

Figure 4: The first 100 components of the 1024 dimension embedding vector for a 30 second clip of a podcast containing both speech and music.

The output scores of the YAMNet have not been calibrated between the different classes, so they cannot be used directly as probabilities. Instead, for a specific task, one probably needs to perform calibration on the classes to get the proper scaling and thresholding of the classes. An example of the YAMNet scores can be found in Figure 5.

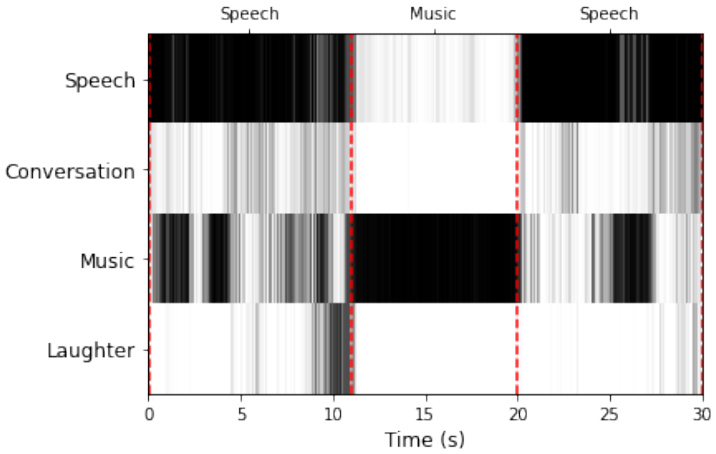

Figure 5: The lograrithm of the scores for the "Speech", "Conversation", "Music" and "Laughter" categories of the AudioSet ontology. The scores are calculated using YAMNet on a podcasts with segments of both speech and music.

Also, since the data is trained on YouTube video clips, if the task data is very different, there might be a mismatch between the two datasets. Besides that, since the model is trained on 10 second AudioSet clips, it might behave differently on events that happen on a smaller timescale.

The YAMNet features are particularly useful because they are very interpretable. The audio

event labels translate directly to sounds we can hear and recognise as humans. Therefore, using these labels, a segment with a particular audio event can be found easily without having to think about the characteristics of the audio event.

2.3 Extracted Features

Using the methods described in the previous sections, we have extracted features for all the episodes in the podcast dataset.

Using the openSMILE toolkit for the eGeMAPS features, we have calculated the functionals for each of the podcasts. In our implementation, the functionals are aggregations over the LLDs of the eGeMAPS over a time window of 1.01 seconds[5] and the features are calculated every 0.48 seconds. In total, there are 88 functionals in the eGeMAPS.

The computation of the eGeMAPS features for the complete dataset took about 5500 hours on a single CPU core. However, the processing was sped up by running the extraction process in parallel on multiple CPU cores.

The resulting feature set is saved as 16-bit floats and compressed into an HDF5 format to reduce the size of the array. The total file size of the eGeMAPS features for all the podcasts is about 75 Gigabytes which is about 4% of the size of the audio data.

Using the pre-trained YAMNet model, we have extracted the 1024-dimension embeddings and the audio event class scores from the raw podcast audio. The YAMNet uses GPU acceleration for the processing of the podcasts audio. Using a single GPU the processing of all the podcasts in the dataset took about 2500 hours, however it was sped up by processing in parallel on multiple GPUs.

Both the embeddings and the audio event class scores from all the podcasts are saved as 16-bit floats and compressed into an HDF5 format. The total size of all the embeddings is about 400 Gigabytes, which corresponds to 20% of the original audio size. The class scores have a total size of 60 Gigabytes for the entire dataset, which corresponds to about 3% of the original size.

All the extracted features presented here will be available with the upcoming TREC 2021 podcast track at `https://podcastsdataset.byspotify.com/`. Also, all the code used to extract the features is available on our GitHub page: `https://github.com/ucl-dis-spotify-group-project/podcast-dataset`.

[5] The implementation was set up to provide a time window which is as close to the 0.96 second windows of the YAMNet features.

3 Metric Engineering

3.1 Creating Labelled Data

In order to devise mood-based metrics, labelled audio data was required. To create the labelled data, mood labels were manually assigned to a sample of 200 2-minute length podcast segments. The sample of segments was selected using the retrieval algorithm Elasticsearch (introduced in Section 4.1) to search the podcast transcripts for specific trigger phrases such that returned segments might be enriched in certain moods. For example, in order to increase the probability of finding funny segments, a search for the phrase "that's so funny" was performed. For finding segments that were likely to be a discussion we searched for the phrase "I agree/disagree". The labelling was performed by the four primary authors of this report, by listening to 50 segments each and noting down any of the relevant labels detailed in Table 3. As seen in the table, many possible labels were considered at this initial stage, and labels were assigned if an instance of a mood occurred at any given time during the two-minute segment.

Mood Category	Label	Definition	No. Segments
Entertaining	funny	funny, or supposed to be funny	103
	storytelling	someone is telling a story	111
	excitement	someone is excited about something being talked about	35
	angry	someone is angry at something being talked about	9
	sad	someone is sad about something being talked about	3
Discussion	narration/monologue	a part with no discussion, just narration/monologue	81
	conversation	a part with a conversation (chit chat, two people or more actively in conversation)	140
	interview	a part with an interview style conversation (more one-sided, someone asks a question and someone answers in monologue)	30
	debate	a part where people debate about something (opinions being voiced)	35
Subjective	approval	someone shows clearly voiced approval of something (e.g. I like X, I love Y)	76
	disapproval	someone show clearly voiced disapproval of something (e.g. I don't like X, I hate Y)	30

Table 3: Table detailing the range of mood labels, their definitions and the number of podcast segments within the sample of 200 segments that had these labels assigned. The sample of segments was selected using Elasticsearch to search the podcast transcripts for specific trigger phrases such that returned segments might be enriched in certain moods. Note that none of the labels are mutually exclusive. Also note that only some of the moods here were subsequently used in analyses.

3.2 Creating Mood Metrics

The mood metrics were subsequently constructed using the labelled data and their corresponding audio features as training data. Note that due to the small size of this data, traditional machine learning techniques could not be employed without a significant risk of over-training. Instead, a more exploratory approach was employed on a case-by-case basis to establish preliminary, "proof of concept" mood metrics. Thus, the aim here is to simply show that it is possible to use the audio data to gain mood-based insights and improve the search task. However, the analysis will be limited, and the results and metrics may need subsequent improvements.

For each mood category (Entertaining, Discussion and Subjective), one specific label was chosen and explored for now. These were funny, debate and disapproval respectively.

Entertaining: Funny metric

To find funny podcast segments, we looked at the "Laughter" feature in the YAMNet data. By looking at how often the "Laughter" score is the highest (or second-highest to only "Speech") for a time step in the data, we can find the amount of laughter in a podcast. Here we excluded the "Speech" category because the "Speech" dominates the YAMNet scores almost always, and we are also looking for segments where people are both laughing and speaking simultaneously. Figure 6 shows the distribution of this feature for both the funny and not funny segments in our labeled set. The figure indicates that if a segment has parts where "Laughter" gets the highest score, it is more likely to be funny than not. Using this threshold of the score being greater than or equal to one to classify our labels results in predicting 68% of the labels in the dataset correctly (for the funny label).

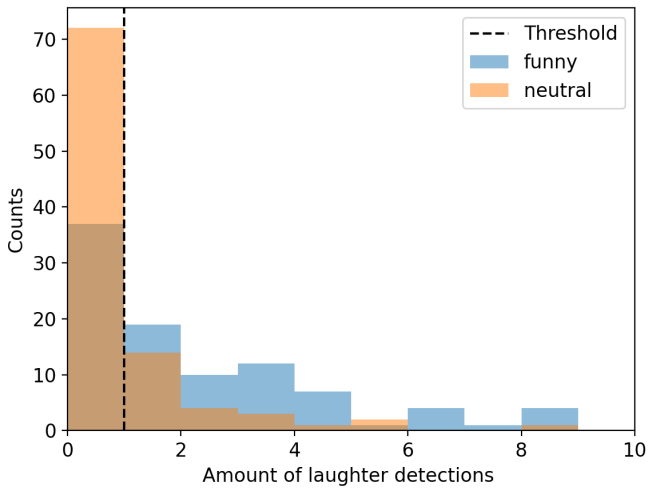

Figure 6: Histogram of how often "Laughter" is the highest score in the YAMNet feature set (excluding "Speech") for segments that are either funny or not funny. We draw a threshold to classify segments as funny when the score is greater than or equal to 1.

Discussion: Debate metric

The Opensmile eGeMAPS features were used to investigate the discussion and subjective moods, where it is predicted their presence will be more uniform across the podcast segment. The mean, standard deviation, maximum and minimum were computed for each eGeMAPS feature across the range of the two-minute segments. Then, the Pearson correlation coefficients were computed for each of these features with their corresponding "debate scores" (score=1 if labelled as debate and 0 otherwise) so that discriminating features could be roughly identified. Using a trial and error approach, a metric for debate was subsequently hand-crafted by linearly combining the features with the largest correlations. This resultant metric is given by Equation 1, and yielded a correlation of 0.39 with the debate score.

$$\begin{aligned} \text{debate metric} = {} & std\_dev(\text{MFCC4\_SMA3\_STDDEVNORM})/143 \\ & + 12 \times max(\text{SLOPEUV500\_1500\_SMA3NZ\_AMEAN})/0.0156 \end{aligned} \tag{1}$$

The discriminating power of this metric can be illustrated by Figure 7. A selection on debate segments can be performed by rejecting all segments below a threshold of debate_metric = 15.

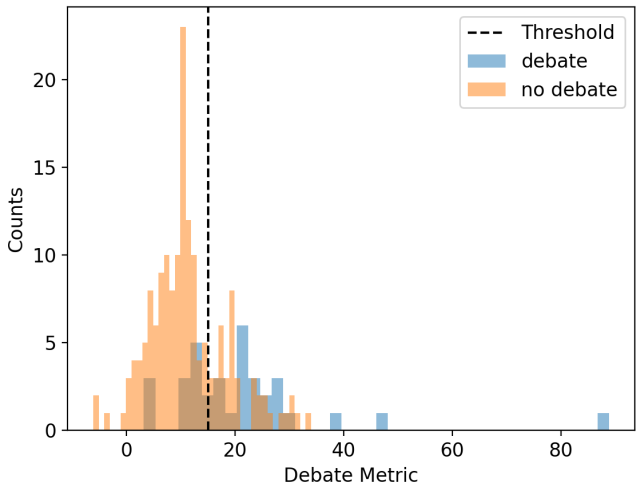

Figure 7: Histogram of the debate metric with its chosen cut value to illustrate its discriminating power.

Subjective: Disapproval metric

An identical approach to the determination of the above debate metric was followed for the disapproval metric. The resultant metric is given by equation 2 and yielded a correlation of 0.29 with the disapproval score.

$$\begin{aligned} \text{disapproval metric} = {} & 2 \times mean(\text{SPECTRALFLUX\_SMA3\_STDDEVNORM})/0.824 \\ & + mean(\text{F1FREQUENCY\_SMA3NZ\_AMEAN})/556 \\ & + mean(\text{F2FREQUENCY\_SMA3NZ\_AMEAN})/1590 \end{aligned} \tag{2}$$

The discriminating power of this metric is illustrated by Figure 8, and a selection on disapproval segments can be performed by rejecting all segments below a threshold of disapproval_metric = 4.2.

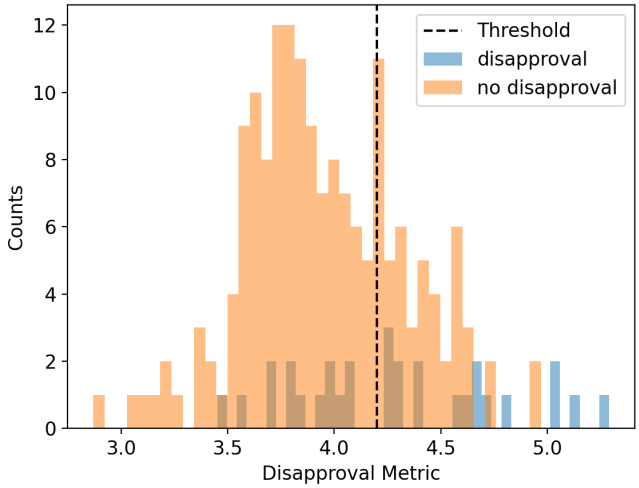

Figure 8: Histogram of the disapproval metric with its chosen cut value to illustrate its discriminating power.

3.3 Performance and Limitations

The classification performance scores were computed for each metric and are detailed in Table 4. In general, all metrics obtain a decent accuracy, and the funny metric also maintains decent recall and precision. However, the debate and disapproval metrics yielded low precision scores due to the significant amount of false positives associated with their cuts. This may be a potential limitation of these metrics.

Mood Metric	Accuracy ($\frac{TP+TN}{Total}$)	Recall ($\frac{TP}{TP+FN}$)	Precision ($\frac{TP}{TP+FP}$)
Funny	0.690	0.641	0.725
Debate	0.745	0.686	0.375
Disapproval	0.700	0.567	0.266

Table 4: Table of classification performance scores for each of the mood metrics. These scores have been computed for the labelled training data only. Note that TP, TN, FP, FN are the true positives, true negatives, false positives and false negatives respectively.

We must also adequately acknowledge the other limitations of our approach. Firstly, the comparatively small size of the labelled data (200 segments) means our metrics are likely to be limited in accuracy and may exhibit over-training. Our predictions are heavily biased by this training dataset and may not do so well on new data. Secondly, whilst we aimed to be consistent in our label definitions, there are invariably some differences in how each individual assigned labels to the segments due to the subjective nature of the moods. Inconsistent labelling in the data set will further limit the accuracy of the devised metrics. Finally, the granularity of the labelling is

a potential concern. We assigned labels to the entirety of the 2 minute segment, but the moods under question may only feature in a fraction of that time, and this will be reflected in the segment features accordingly. We may therefore miss distinguishing features when computing averages across the 2 minute range, and this will limit the performance of the metrics.

4 Segment Retrieval

Here, we present our implementation for the updated segment retrieval task. First, we outline our baseline segment retrieval procedure using exclusively the textual features of each segment matched to the topical title and description for the query. We then present our methodology that builds upon this baseline using the audio features of each segment matched to the mood of the query. The combination of these procedures is outlined in Figure 9, which summarises our full segment retrieval implementation.

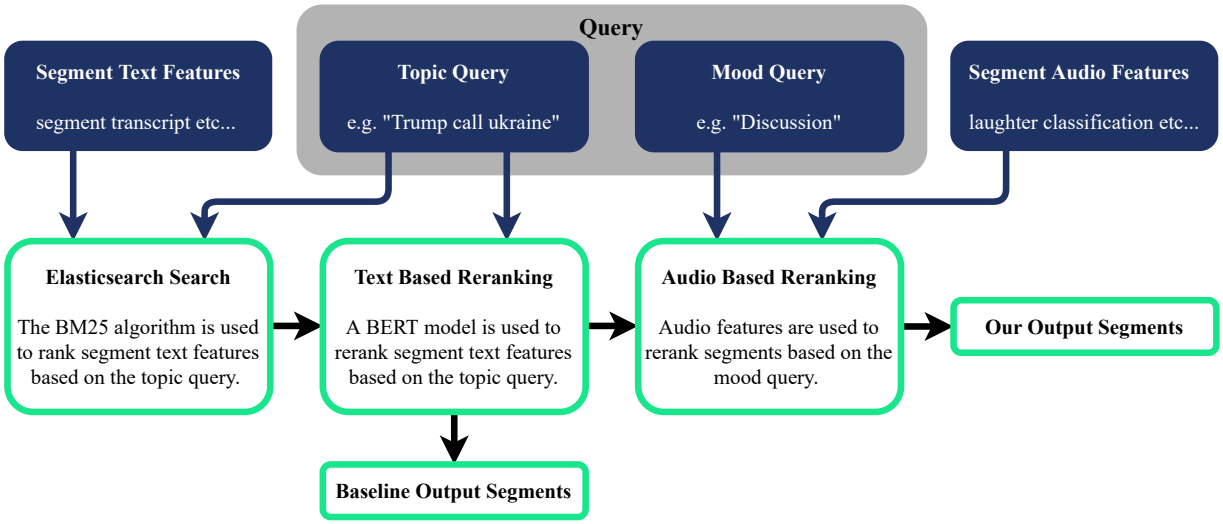

Figure 9: Diagram of our complete segment retrieval procedure for a given query containing both a topic (title and description) and a mood.

4.1 Baseline retrieval

For the first stage of the baseline retrieval system, we used the standard BM25 retrieval algorithm [10] built into Elasticsearch[6]. For this purpose, we created a single document index containing all 3.4 million two-minute-long podcast segments, each with three fields: the episode name, the episode description, and the individual segment transcript. We then used the concatenation of the topical query title and description to perform a multi-match search of the index querying all three fields at once. As part of the query, the segment transcript field score was multiplied by two as we deemed this field the most important for determining the topical

[6] `https://www.elastic.co/elasticsearch`

relevancy of the segment. For each query, we then returned the highest 100 scoring segments from Elasticsearch.

For the next stage of the baseline retrieval system, we used a BERT-based [2] text reranking model[7]. Such models are currently considered the standard implementation for achieving near state-of-the-art performance. Specifically, we used a BERT base uncased model that had been fine-tuned on the MS MARCO dataset [9] and employed via the Huggingface transformers library [11]. The MS MARCO dataset contains 400 million tuples of a query and relevant and non-relevant passages, such that the fine-tuned model should learn topical relevancy mappings.

For each query and passage fed into the reranking model as sentence A and sentence B, with BERT tokenisation, the model outputs a score (between -10 and 10) estimating how relevant the passage is to the query. In our case, we concatenated the topical title and description from the query as sentence A and used the concatenation of the segment transcript, episode name, and episode description for sentence B. The full text was then truncated to fit within the maximum length of 512 tokens. The model was evaluated on each segment returned from Elasticsearch in conjunction with the query, and the segments were then reranked using the output score exclusively to return the final baseline output segments.

4.2 Audio-enhanced retrieval

Using the audio derived metrics previously outlined in Section 3 we then built upon our baseline retrieval system to help categorise the segments as entertaining, subjective, or discussion. For each of the 100 segments retrieved by the baseline system (regardless of Elasticsearch or reranking model score), this involved applying a series of selection cuts for each mood which are given below.

For the entertaining mood, segments were selected by requiring that:

- The YAMNet derived funny metric was greater than or equal to 1.0.

For the subjective mood, segments were selected by requiring that:

- The eGeMAPS derived disapproval metric was greater than 4.2.

For the discussion mood, segments were selected by requiring that:

- The eGeMAPS derived debate metric was greater than 15.0; and
- The number of YAMNet frames where the "Narration" label was greater than 0.02 was less than 100.

We additionally required that the number of YAMNet frames where the "Music" label was greater than 0.02 was less than 100 for all moods, as the audio derived metrics within segments with lots of music were found to be heavily skewed from the bulk. We then returned up to 10 selected segments for each mood category ranked by their topical reranking model output score from the baseline implementation for relevancy.

[7] `https://huggingface.co/amberoad/bert-multilingual-passage-reranking-msmarco`

4.3 Retrieval results and discussion

For each of the topical query titles and descriptions in Table 2 we ran the entire segment retrieval pipeline, returning the top 10 segments from the baseline (text-based) system, as well as the top 10 segments for each possible mood from the audio-enhanced mechanism (sometimes less than 10 if a smaller number were selected by a mood selection).

Using the same procedure as in Section 3.1, the four primary authors of this report then assigned the labels from Table 3 to all the selected segments. Beforehand, all the segments were randomised to reduce labelling bias. We then classified entertaining segments as those which we labelled as "funny", subjective segments as those which we labelled as either "approval" or "disapproval", and discussion segments as those which we labelled as either "debate" or "conversation".

Table 5 details how many of the segments for both the baseline and audio-enhanced selections were classified as each mood for each query and mood combination. When the results are combined across all queries, the baseline implementation contains 20% entertaining segments while the audio-enhanced selection achieves 77%, 23% discussion segments while the audio-enhanced selection achieves 43%, and 37% subjective segments while the audio-enhanced selection achieves 44%.

Topical query title	Query mood	Baseline (text)	Ours (audio-enhanced)
Near death experiences	Entertaining	0/10	6/6
	Discussion	0/10	4/10
	Subjective	3/10	4/10
Black lives matter	Entertaining	3/10	7/10
	Discussion	5/10	8/10
	Subjective	5/10	5/10
Halloween stories and chat	Entertaining	3/10	7/10
	Discussion	2/10	1/10
	Subjective	3/10	2/5

Table 5: The number of segments classified as each mood by both the baseline (text-based) retrieval system and our audio-enhanced retrieval mechanism for three test queries. The numbers are given as fractions out of the total number of segments considered; this is as the audio-enhanced mechanism found less than ten appropriate segments in some cases.

Our results clearly show that the audio features can be used to enhance the retrieval of podcast segments when an additional mood requirement is added to the query. This improvement is found to be particularly true for the classification of entertaining segments, where our use of the YAMNet derived "laughter" metric is found to increase the fraction of entertaining segments returned by nearly four times.

Although we did not consider how the text features may be used to classify the mood of segments, which may prove significant, it is promising to see some signal within the audio for such a task exists. It is highly likely that future work conducting a more thorough exploration of such audio metrics, and using a greater amount of labelled data, would lead to much-improved performance, particularly in the discussion and subjective mood categories.

It must also be noted that we did not study how the topical relevancy of the segments was affected by applying the audio-enhanced selections. Some form of trade-off will need to be decided upon to achieve both a relatively topical relevant segment at the same time as being mood appropriate. Again we leave this as future work.

5 Conclusion

We have extracted two sets of high-level features across the full Spotify Podcasts Dataset and made them readily available to others. Not only does this mean that others will have access to useful audio features, but also they will not have to deal with large data files or repeat the computationally intensive extraction. The extracted features include the 88 eGeMAPS functionals, the 1024-dimensional YAMNet embedding, and the YAMNet audio event class scores.

By labelling a small subset of the podcast dataset, we derived three audio metrics from the high-level features to classify segments as funny, consisting of a debate, or containing disapproval. Although only a simple first attempt, we achieved a classification accuracy of 69%, 75%, and 70% for each metric, respectively, proving that such audio metrics are possible.

We implemented a baseline segment retrieval system using an Elasticsearch index and a BERT-based relevancy reranking model. On top of this baseline, we used the three audio metrics derived from the YAMNet and eGeMAPS features to implement an audio-enhanced retrieval mechanism to improve the selection of segments as either entertaining, subjective or containing a discussion. Impressively, we increased the fraction of entertaining segments by nearly four times compared to the baseline by using the YAMNet derived "funny" metric.

Using the features and methods discussed in this report, we also hope to make a submission to the 2021 TREC podcast track tasks.

6 Future Work

To improve upon the presented results, a first step would be to enhance and expand the labelled dataset. While we labelled 2-minute segments, the moods we were trying to characterise often spanned much smaller time frames. Therefore, an improved set would contain labels for shorter length podcast segments. Also, because of the sparsity of the data and the selection process, we found it challenging to find good metrics for each of the moods. Providing a lot more data will allow for the training of more complex and accurate models.

Another possible way the results could be improved is by training a model on the 1024 dimension embedding vectors from the YAMNet feature extraction. Using only a small labelled dataset, these embeddings could prove helpful in detecting specific audio events of interest.

Finally, the results from our mood metrics could be combined with any number of text-based features extracted from the segment transcripts. By combining the information from the transcript and the audio, new, improved metrics might be found.

Bibliography

[1] Ann Clifton et al. "100,000 Podcasts: A Spoken English Document Corpus". In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020, pp. 5903–5917.

[2] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

[3] F. Eyben et al. "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing". In: *IEEE Transactions on Affective Computing* 7.2 (Apr. 2016), pp. 190–202. ISSN: 1949-3045. DOI: 10.1109/TAFFC.2015.2457417.

[4] Florian Eyben, Martin Wöllmer and Björn Schuller. "Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor". en. In: *Proceedings of the International Conference on Multimedia - MM '10*. Firenze, Italy: ACM Press, 2010, p. 1459. ISBN: 978-1-60558-933-6. DOI: 10.1145/1873951.1874246.

[5] Jort F. Gemmeke et al. "Audio Set: An Ontology and Human-Labeled Dataset for Audio Events". en. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA: IEEE, Mar. 2017, pp. 776–780. ISBN: 978-1-5090-4117-6. DOI: 10.1109/ICASSP.2017.7952261.

[6] Andrew G. Howard et al. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications". In: *arXiv:1704.04861 [cs]* (Apr. 2017). arXiv: 1704.04861 [cs].

[7] Rosie Jones et al. "Current Challenges and Future Directions in Podcast Information Access". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021.

[8] Rosie Jones et al. "TREC 2020 Podcasts Track Overview". In: *Proceedings of the Twenty-Ninth Text REtrieval Conference (TREC 2020)*. Ed. by Ellen M. Voorhees and Angela Ellis. Gaithersburg: NIST, 2021.

[9] Tri Nguyen et al. "MS MARCO: A human generated machine reading comprehension dataset". In: *CoCo@ NIPS*. 2016.

[10] Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[11] Thomas Wolf et al. "HuggingFace's Transformers: State-of-the-art natural language processing". In: *arXiv preprint arXiv:1910.03771* (2019).