Audio Features, Precomputed for Podcast Retrieval and Information Access Experiments

Abigail Alexander[1] Matthijs Mars[1] Josh C Tingey[1] Haoyue Yu[1]
Chris Backhouse[1] Sravana Reddy[2]
Jussi Karlgren[2]

[1] University College London, London, United Kingdom

[2] Spotify, Boston, United States and Stockholm, Sweden

This paper has been published in the Proceedings of the CLEF Conference

Abigail Alexander, Matthijs Mars, Josh Tingey, Haoyue Yu, Chris Backhouse, Sravana Reddy, and Jussi Karlgren 2021. "Audio Features, Precomputed for Podcast Retrieval and Information Access Experiments". In *Proceedings from the International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer Lecture Notes in Computer Science. DOI: 10.1007/978-3-030-85251-1_1 (Cf. also a longer tech report by Alexander et al. (2021) with more technical detail.)

Abstract

This paper describes how an existing collection of podcast material has been enriched with precomputed audio features. The feature set which is described in the paper is made available to facilitate more convenient information access experimentation to collections that include both audio and text data. A simple example analysis is given to demonstrate how the audio features can be used to score podcast segments for being entertaining, discussion oriented, or subjective, to fit the current TREC Podcast Track task.

1 Podcasts are a new medium

Access to podcasts involves new research and development challenges. Most of the starting points of current retrieval technology for both speech and text take departure in (1) topical search being the primary access path to a collection and (2) in transcribed text being the most convenient way to represent the content of spoken material. This is arguably true for most task-oriented information access use cases, but for entertainment and enjoyment, use cases which are at the forefront of attention for podcast listeners, the ranking criteria for candidate items are likely to be broader than topical relevance. (Jones et al. 2021b)

Speech, which in general is less conventionalised and less rule-bounded than writing, is a richer communicative channel than text. Podcast material is different from previous spoken language material in

several ways — the production circumstances, the intended use cases, and the distribution technology conspire to render the language in podcasts different from known related genres such as radio broadcasts, recorded lectures, conversations, or written interactive usage such as chats or forum discussions.

Some podcasts track closely to genre conventions known from text or previous practice, where others are more conversational with rapid exchanges of ideas, quick conversational moves, argumentation, and overlapping speech. Podcasts can be monologues, lectures, conversations, interviews, debates, and chatty multi-party conversations. They may contain historical clips, and the range of emotions expressed by participants can range across most human sentiments: in fact, one of the apparent attractions of podcasts as a medium is that the constraints of other media can be and often are breached with impunity by the participants in a podcast conversation.

Transcriptions normalise much of this type of variation since text is designed to express emotions in explicit terms rather than intonation and to render topical content in an orderly way. Most transcription technologies will handle linguistic variation — dialectal and sociolectal variation, e.g. — badly and will fail entirely in the face of e.g. rapid multi-party discourse, overlapping speech, disfluency and repair, or ambient noise. These various variational dimensions are not irrelevant to listeners, however: listeners do pay attention to various auditive characteristics of a podcast when they assess the qualities of an episode or a show. Martikainen (2020)

This means that information access technology for podcast material cannot entirely be based on current text-based retrieval technology. Retrieval systems and exploration tools for podcasts must in some way take into account such information that would be lost when transcribing the audio content to text. This paper presents a set of precomputed audio features to lower the threshold to systematic experimentation on information access—retrieval, clustering, classification, summarisation, and related applications—for podcast material.

2 The Spotify English-language Podcast Dataset

In 2020, Spotify released a large dataset of podcasts[1] with the view of enabling researchers from various fields to test their theories and hypotheses on realistic scale collections of spoken audio, particularly podcast material. The dataset consists of 105,360 English language podcast episodes collected from Spotify's catalogue between late 2019 and early 2020. Each episode includes the audio (sampled at 44.1 kHz), an automatically generated transcript, and some associated metadata, including the episode and show names and descriptions. 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. Clifton et al. (2020)

3 The TREC Podcasts Track

For this purpose, TREC, the annual Text Retrieval Conference, in 2020 organised a Podcasts Track with two information access tasks to experiment on podcast material Jones et al. (2021a). There have been previous speech retrieval tracks both in TREC and in CLEF, but this is the first initiative to address the specifics of podcast material, recognising the challenges outlined above.

The Podcasts Track attracted great interest but not all registered participants submitted results and no participant made direct use of the audio data, except to provide alternative transcriptions. A post-participation questionnaire indicated that the size of the collection and technical challenges with audio analysis were considerable thresholds for participants. The objective for the second year is to lower participation thresholds in general and to specifically provide the participants with precomputed audio features to encourage experimentation on the audio data rather than the transcripts.

3.1 Task 1: Fixed-length Segment Retrieval

The segment retrieval task asks participants to, given a query, retrieve appropriate two-minute segments from the data set. In 2020, the queries were of three types: *topical*, *known-item*, and *refinding* queries, all based on topical relevance as the primary target criterion. These were all addressable using text retrieval

[1] `https://podcastsdataset.byspotify.com/`

techniques on the transcripts, and the participants' approaches were well aligned, using text retrieval, reranked using deep learning models.

In 2021, to encourage participants to use the audio material, the task types have been modified. The refinding queries and the known-item queries, which overlapped to some extent, have been combined into one type. The topical queries will be assessed differently: the participants will be asked to submit results for the topical queries separately ranked by several different target notions: as before, by *relevance*, but also, separately ranked by if the segment is *entertaining*, such that the segment presents the topic in a way which the speakers intend to be amusing and entertaining to the listener, if the segment is *subjective*, such that the speakers explicitly and clearly make their approval or disapproval of the topic evident, and if the segment contains *discussion* with more than one speaker contributing to the topic. In addition, a *speaker* query type is added, for which the participants will be asked to find episodes where some given speakers participate. For speaker queries, a clip of the intended speaker taken from another recording will be given for reference.

We expect that these reranking criteria will be better served by making use of the audio data in addition to the topical retrieval (which presumably will continue best being executed by using the transcript).

3.2 Task 2: Summarisation

The summarisation task asks of participants to, given a podcast episode, its audio, and transcription, return a short text snippet capturing the most important information in the content. Returned summaries should be grammatical, standalone utterances of significantly shorter length than the input episode description. The user task is to provide a short description of the podcast episode to help the user decide whether to listen to a podcast. In 2021, the participants will be asked to provide, in addition to the textual summary, up to three audio clips from the podcast to give the user a sense of what the podcast sounds like. The audio clips will be assessed by human assessors to answer the question "Do the clips give a sense of what the podcast sounds like, (as far as you can tell from listening to it)?"

We expect that this task will benefit greatly from using audio features.

We look forward to seeing how these tasks will be addressed by participants using audio features alongside textual features, and expect to see various levels of utility for the precomputed features. We expect that in coming years, the feature sets and tools to extract them from the speech signal will evolve in capacity and effectiveness for downstream tasks. This effort is a step in that direction: by sharing precomputed audio features for the English-language Podcast Data Set suitable for the TREC tasks, we hope to see more experimentation on audio as well as new features and audio analyses be made available.

4 Feature Extraction

As all the podcasts in the English-language Podcast Data Set are sampled at 44.1 kHz, a 30-minute long podcast contains approximately 80 million data points. This is illustrated in Figure 1 which shows an example of the 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. In order to extract useful information from the podcast audio, the data needs to be processed into more informative high-level features that act to summarise the raw audio signal.

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, employing the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) Eyben et al. (2016), uses established features in phonetics and speech sciences. The second uses the learned features of the Yet Another MobileNet (YAMNet)[2] model to label the data with labels from the AudioSet ontology Gemmeke et al. (2017).

4.1 The Geneva Minimalistic Acoustic Parameter Set

The Geneva Minimalistic Acoustic Parameter Set or GeMAPS is an attempt by Eyben et al. Eyben et al. (2016) to design a minimalistic and standardised parameter set of acoustic features that are useful for

[2] `https://github.com/tensorflow/models/tree/master/research/audioset/yamnet`

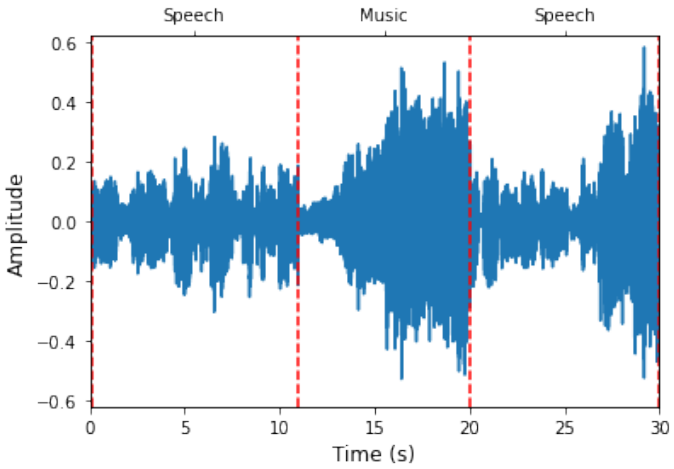

Figure 1: The audio waveform of a 30-second clip from a podcast containing sections of both speech and music.

machine learning problems. These features include parameters 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). Importantly, all are calculated in a clearly defined, standardised manner, such that the values are reproducible and results can be easily compared.

The feature set is designed to be minimalistic in that it calculates the least amount of features required to generate relatively strong results. This way, classifiers trained on the features are less likely to over-adapt to the training data and instead generalise well. The interpretation of the parameters and results from a minimalistic set is also more straightforward since the features and derived models are relatively simple.

The minimal set contains 18 Low Level Descriptors (LLDs) describing vocal features such as intonation, stress, rhythm, excitation, as well as various spectral descriptors that analyse the base frequency and harmonics of speech. The selected LLDs are chosen based on their relative importance from previous research results. This minimal set is referred to as GeMAPS. In addition, an extension set with seven further LLDs, all cepstral descriptors which analyse the periodic structures in frequency data Devlin et al. (2019), is also defined. 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-force 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 some predictive tasks. Classification with eGeMAPS achieves similar (if somewhat reduced) performance to the large scale parameter sets, yet the size of the parameter set is only 2% of the most extensive set included in the comparison Eyben et al. (2016).

The eGeMAPS features can be computed from the raw audio waveform using the openSMILE feature extraction toolkit Eyben et al. (2010). 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 [3]. Figure 2 shows a subset of the eGeMAPS features for the same podcast segment as shown before in Figure 1.

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

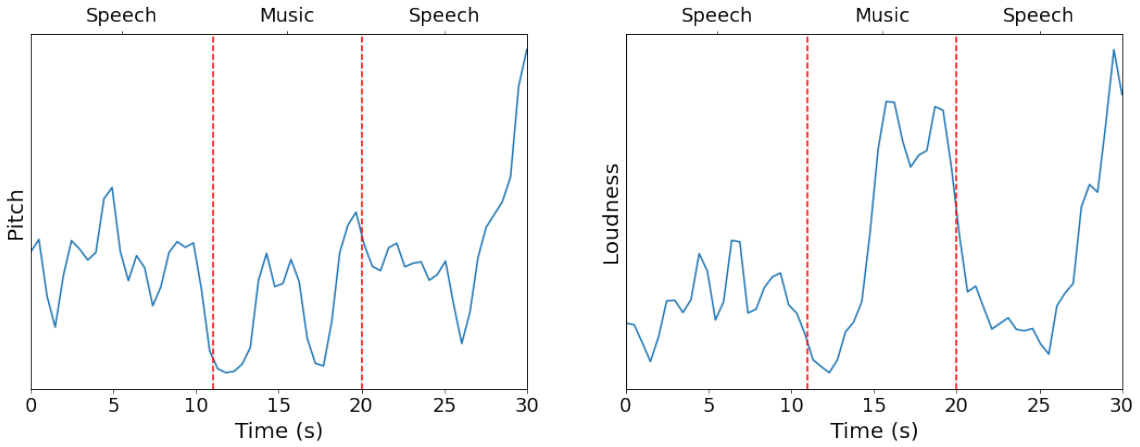

Figure 2: The pitch and loudness averaged over 0.96 seconds of audio for a clip of 30s from the Podcast Dataset. The data is sampled every 0.48 seconds.

4.2 YAMNet

The second feature set uses learned features that are inferred from a labelled dataset. Yet Another MobileNet or YAMNet is a neural network based audio feature extractor based on the Mobilenet_v1 convolutional architecture Howard et al. (2017), which is trained on the AudioSet corpus. AudioSet Gemmeke et al. (2017) contains audio from approximately 2 million 10-second YouTube clips, labelled by humans into 521 audio event classes. The complete AudioSet ontology covers sound classes such as humans, animals, and music, amongst many other common everyday environmental sounds.

The network pre-processes audio input into mel spectrograms—spectrograms where the frequency spectrum is transformed to fit human perception—and uses a convolutional neural network (CNN) to analyse these spectrograms in a similar manner to image recognition tasks.

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

Besides using the AudioSet labels from YAMNet as an audio event classifier, the 1024-dimensional embedding vector can be used as a general-purpose audio feature representation. The pre-trained YAMNet model can then be used as a feature extractor for a smaller network trained on top of the embedding vector on a small set of labelled data for a particular task without retraining the complete network from scratch. Figure 3 shows the vectors for a podcast segment from the Podcast Dataset.

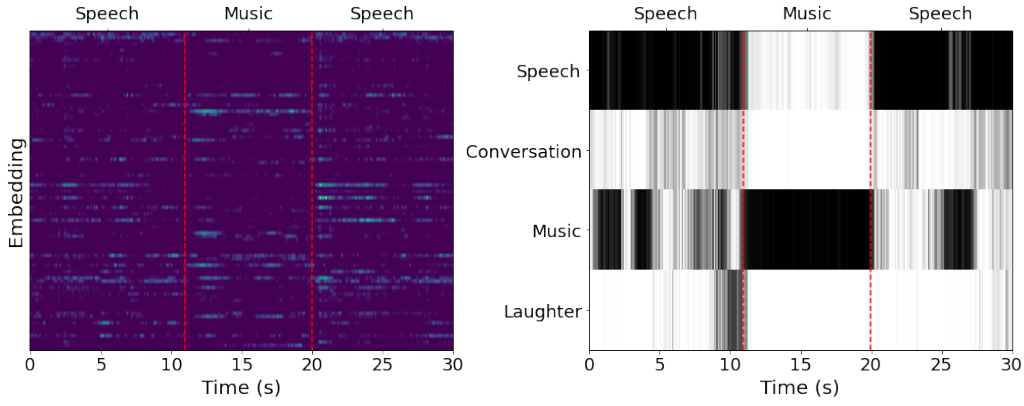

Figure 3: *(Left)* The first 100 components of the 1024 dimensional vector for a 30 second clip of a podcast containing both speech and music. *(Right)* The logarithm of the class scores for the event labels speech, conversation, music, and laughter for the same podcast segment.

The output scores of YAMNet are not calibrated between the different classes, so they cannot be used directly as probabilities. Instead, for a specific task, one needs to perform calibration across the classes to determine the appropriate scaling and thresholding of the classes.

Since the data is trained on YouTube video clips using 10-second AudioSet clips, there is a risk of mismatch if the task data are different or if the events of interest occur 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 easily found without considering its underlying acoustic characteristics (Martikainen 2020, cf).

4.3 Extracted Features

We have extracted both above sets of features for every episode in the Podcast Dataset.

Using the openSMILE toolkit, we have calculated each podcast's eGeMAPS functionals. In our implementation, the functionals are aggregations (mean, standard deviation, etc...) of the eGeMAPS LLDs over a time window of 1.01 seconds[4] starting every 0.48 seconds, such that all the windows overlap both their neighbours by approximately half their length. In total, there are 88 functionals in the eGeMAPS feature set. The computation of the eGeMAPS features for the complete Podcast Dataset would take $\sim$ 5500 hours on a single CPU core. Therefore, the processing was sped up by running the extraction process in parallel on multiple CPU cores. The resulting eGeMAPS feature set is saved as 16-bit floats and compressed into an HDF5 format to reduce the storage size. The resulting total file storage size of the eGeMAPS features for all the podcasts is approximately 75 Gigabytes, which is $\sim$ 4% the size of the raw audio data.

Using the pre-trained YAMNet model, we extracted the 1024-dimensional embedding vectors and the audio event class scores from the podcast audio. These are calculated for every 0.96 second long window of the podcast starting every 0.48 seconds, such that all the windows overlap both their neighbours by approximately half their length. GPU acceleration was used to speed up the processing of the podcast audio. Using a single GPU, the processing of all the podcasts in the Podcast Dataset would take $\sim$ 2500 hours; therefore, it was sped up by processing in parallel on multiple GPUs. The 1024-dimensional embedding vectors 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 vectors is approximately 400 Gigabytes, which corresponds to $\sim$ 20% of the original audio size. The class scores have a total size of 60 Gigabytes for the entire Podcast Dataset, which corresponds to $\sim$ 3% of the original audio size.

5 Example Analysis: TREC 2021 reranking tasks

5.1 Creating Labelled Data

In order to devise some example mood-based metrics to fit the target notions for the TREC tasks, labelled audio data are required. To create the labelled data, mood labels were manually assigned to a sample of 200 2-minute-long podcast segments. To enrich the sample used, we selected the segments based on a search of an Elasticsearch[5] index containing all the possible Podcast Dataset segment transcripts. For example, to find funny segments, the phrase "that's so funny" was used as the query; for subjective segments, the phrases "I agree" and "I disagree". The authors performed the labelling by listening to a subset of segments each and noting down any of the relevant labels detailed in Table 1. A label was assigned if an expression of a mood occurred at any given time during the two-minute segment; multiple labels for a segment were allowed.

5.2 Creating Mood Metrics

The mood metrics were subsequently manually formulated using the labelled data and their corresponding audio feature scores. Due to the small size of the data set, we did not employ traditional machine learning

[4]This window length was chosen to provide a time window which is as close as possible to the 0.96 second windows of the YAMNet features.

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

Table 1: Table detailing the range of mood labels, their definitions, and the number of podcast segments out of the 200 segment sample that had these labels assigned. The labels are not mutually exclusive.

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	35
	angry	someone is angry at something	9
	sad	someone is sad about something	3
Discussion	narration/monologue	segment with no discussion	81
	conversation	segment with conversation (chit chat, two people or more actively in conversation)	140
	interview	segment with interview style conversation (more one-sided, someone asks a question and someone answers in monologue)	30
	debate	segment where people debate about something (opinions being voiced)	35
Subjective	approval	clearly voiced approval (e.g. I like X, I love Y)	76
	disapproval	clearly voiced disapproval (e.g. I don't like X, I hate Y)	30

techniques. Instead, a more exploratory approach was employed on a case-by-case basis to establish preliminary, "proof of concept" mood target notion metrics. These results and metrics are not definitive and serve only to demonstrate that it is possible to use the audio data to gain mood-based insights to improve the search task. For each target notion category (Entertaining, Discussion and Subjective), one specific label was chosen and explored: *funny*, *debate*, and *disapproval*, respectively.

5.2.1 Entertaining: Funny metric

To find funny podcast segments, we used the "Laughter" feature from YAMNet. By tabulating how often the "Laughter" score is the highest (or second-highest to only "Speech") for each time step in the data, we can find the amount of laughter in a podcast. Figure 4 shows the distribution of this feature for both the funny and non-funny segments in our labelled 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 a greater than or equal to one threshold for the score to classify our labels results in predicting 69% of the *funny* labels in the Podcast Dataset correctly.

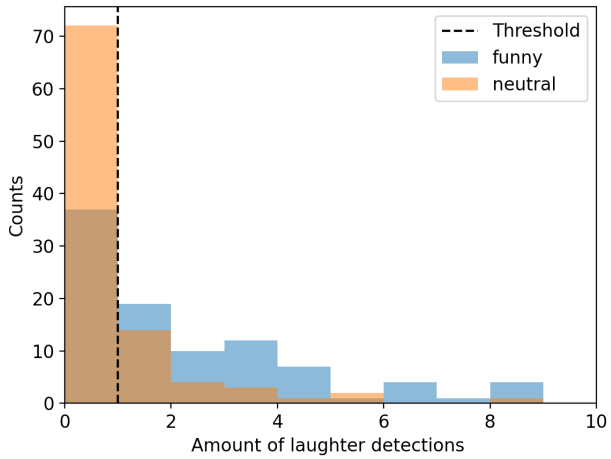

Figure 4: Histogram for the laughter metric with its chosen threshold cutoff value.

5.2.2 Discussion: Debate metric

The Opensmile eGeMAPS features were used to investigate the discussion and subjective notions, 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 sample of 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 most significant correlations. This resultant metric is given by Equation 1 with a distribution as shown in Figure 5. A selection on debate segments can be performed by rejecting all segments below a threshold of debate_metric = 15. Using this threshold, we succeed in predicting 74% of the *debate* labels in the Podcast Dataset correctly.

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

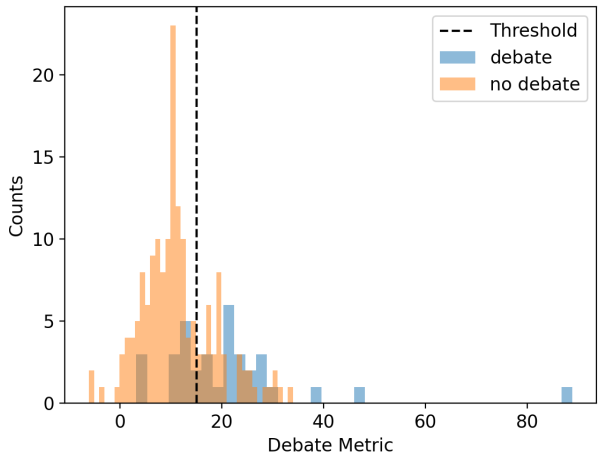

Figure 5: Histogram for the debate metric with its chosen threshold cutoff value.

5.2.3 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.The discriminating power of this metric is illustrated by Figure 6, and a selection on disapproval segments can be performed by rejecting all segments below a threshold of disapproval_metric = 4.2. This cutoff value corresponds to an accuracy of 0.700 on the training data.

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

5.3 Performance and Limitations

The above classification performance scores were computed for each metric and are detailed in Table 2. In general, all metrics yield reasonable accuracy, even for this small and subjectively labelled training set of only 200 segments. The *funny* metric also maintains good recall and precision. However, the *debate* and *disapproval* metrics show low precision scores due to the amount of false positives associated with

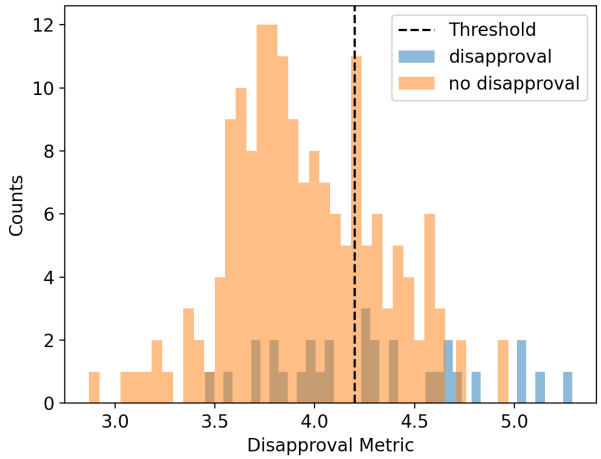

Figure 6: Histogram for the disapproval metric with its chosen threshold cutoff value.

the simple linear cuts. This indicates that more sophisticated classification methods, larger training sets, and possibly more consistent labelling of the training set are necessary. In addition, it is worth to note the likely systematic effect of segment and sentiment granularity mismatch. Here, we assigned labels to the entirety of the 2-minute segment, but the moods in question may only feature in a fraction of that time, and this will be reflected in the segment features accordingly. Averaging over the 2-minute range risks missing distinguishing features.

Table 2: Mood metric classification performance for manually labelled training data.

Mood Metric	Accuracy	Recall	Precision
Funny	0.690	0.641	0.725
Debate	0.745	0.686	0.375
Disapproval	0.700	0.567	0.266

6 Concluding remarks

We wish to encourage more researchers to use audio data for podcast analysis, e.g. in the TREC Podcasts Track. All the extracted features presented here are available in a simple format with the entire Podcast Dataset[6] and the code used to extract the features is available on GitHub[7]. Alexander et al. (2021) We expect to see other audio analyses and feature sets added to further enrich the Podcast Dataset and intend our effort to be a model for how such features and analyses can be shared to facilitate audio-based experimentation on speech at realistic scale.

[6] https://podcastsdataset.byspotify.com/
[7] https://github.com/trecpodcasts/podcast-audio-feature-extraction

References

Abigail Alexander, Matthijs Mars, Josh Tingey, and Haoyue Yu. Audio-enhanced segment retrieval within the spotify podcasts dataset. Technical report, University College London, 2021.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, et al. 100,000 Podcasts: A Spoken English Document Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*. Association for Computational Linguistics, 2019.

F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016. ISSN 1949-3045.

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the International Conference on Multimedia - MM '10*. ACM Press, 2010. ISBN 978-1-60558-933-6.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, 2017. IEEE. ISBN 978-1-5090-4117-6.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv:1704.04861 [cs]*, April 2017.

Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth JF Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. TREC 2020 Podcasts Track Overview. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of the Twenty-Ninth Text REtrieval Conference (TREC)*. NIST, 2021a.

Rosie Jones, Hamed Zamani, Markus Schedl, Ching-Wei Chen, Sravana Reddy, Ann Clifton, Jussi Karlgren, Helia Hashemi, Aasish Pappu, Zahra Nazari, LongQi Yang, Oguz Semerci, Hugues Bouchard, and Ben Carterette. 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*, 2021b.

Katariina Martikainen. Audio-based stylistic characteristics of podcasts for search and recommendation: a user and computational analysis. Master's thesis, University of Twente, 2020.