100,000 Podcasts: A Spoken English Document Corpus

Ann Clifton
Spotify
aclifton@spotify.com

Sravana Reddy
Spotify
sravana@spotify.com

Yongze Yu
Spotify
yongzey@spotify.com

Aasish Pappu
Spotify
aasishp@spotify.com

Rezvaneh Rezapour*
University of Illinois
at Urbana-Champaign
rezapou2@illinois.edu

Hamed Bonab*
University of Massachusetts
Amherst
bonab@cs.umass.edu

Maria Eskevich
CLARIN ERIC
maria@clarin.eu

Gareth J. F. Jones
Dublin City University
gareth.jones@dcu.ie

Jussi Karlgren
Spotify
jkarlgren@spotify.com

Ben Carterette
Spotify
benjaminc@spotify.com

Rosie Jones
Spotify
rjones@spotify.com

Abstract

Podcasts are a large and growing repository of spoken audio. As an audio format, podcasts are more varied in style and production type than broadcast news, contain more genres than typically studied in video data, and are more varied in style and format than previous corpora of conversations. When transcribed with automatic speech recognition they represent a noisy but fascinating collection of documents which can be studied through the lens of natural language processing, information retrieval, and linguistics. Paired with the audio files, they are also a resource for speech processing and the study of paralinguistic, sociolinguistic, and acoustic aspects of the domain. We introduce the Spotify Podcast Dataset, a new corpus of 100,000 podcasts. We demonstrate the complexity of the domain with a case study of two tasks: (1) passage search and (2) summarization. This is orders of magnitude larger than previous speech corpora used for search and summarization. Our results show that the size and variability of this corpus opens up new avenues for research.

1 Introduction

Podcasts come in many formats and levels of formality. Episodes appear on a regular or irregular cadence. They can be formal news journalism or conversational chat; fiction or non-fiction. They are sharply growing in popularity (Whitner, 2020) and yet have been relatively little studied. This medium opens up a rich palette of questions and issues for research in speech and language technology, linguistics, information access, and media studies.

To facilitate research into podcasts, we have produced a corpus of podcast episodes, comprising nearly 60,000 hours of speech. This is orders of magnitude larger than previous transcribed speech datasets, and contains a rich variety of genres, subject matter, speaking styles, and structural formats. Our contributions are four-fold:

- The largest corpus of transcribed speech data, from a new and understudied domain,
- A set of labeled data for retrieval and summarization on this corpus,
- Benchmarking results for retrieval and summarization tasks using standard baselines,
- An analysis of the data and benchmarking results, highlighting domain differences from vanilla versions of these tasks to motivate areas of future research.

The corpus can be accessed at `podcastsdataset.byspotify.com`.

[*] Work done while at Spotify

Proceedings of the 28th International Conference on Computational Linguistics, pages 5903–5917
Barcelona, Spain (Online), December 8-13, 2020

2 Related Datasets

Earlier speech corpora contained relatively clean audio, often with a single speaker reading from a prepared text, such as the TIMIT collection (Garofolo et al., 1990) or broadcast news corpora, which have been used as data sets for speech retrieval experiments in both TREC (Garofolo et al., 2000) and CLEF (Federico and Jones, 2003), and for Topic Detection and Tracking (Allan et al., 1998). These more formal settings or samples of formal content are useful for the study of acoustic qualities of human speech, but represent a more idealized scenario than practical audio processing tasks of interest today.

Conversational datasets with noisier speech have been collected for specific domains, often intended to capture regularities of some particular communication situation, such as the ATIS corpus of air travel information requests (Hemphill et al., 1990), meeting recordings (Garofolo et al., 2004b), telephone conversations (Canavan et al., 1997; Godfrey and Holliman, 1993), and broadcast news (Garofolo et al., 2004a). There are some collections of more naturally occurring conversational material such as the CALLHOME corpus (Canavan et al., 1997), the Santa Barbara Corpus of Spoken American English (Bois and Engebretson, 2005) and the TED talks corpus (Hasebe, 2015). While some of the content in such collections share characteristics with podcast material, podcasts' combination of unscripted and spontaneously organised discourse in a conversational setting, with turntaking, interviews, stretches of monologue, argumentation, and the inclusion of other audio material including non-speech segments is not yet represented in any collection of spoken language available with transcripts for research purposes.

For summarization corpora in particular, the CNN/DailyMail data (Hermann et al., 2015) is one of the few large summarization datasets with manually written summaries. Spoken document summaries are also available for the AMI meeting corpus (Mccowan et al., 2005) and the ICSI meeting corpus (Janin et al., 2003), as well as corpora of lectures (Miller, 2019), and voicemail (Koumpis and Renals, 2005). Spina et al. (2017) collect and evaluate 217 hours of podcasts for query-biased extractive summarization. In recent work, Tardy et al. (2020) train a model to reproduce full-length manual reports aligned with automatic speech recognition transcripts of meetings, and Gholipour Ghalandari et al. (2020) generate a corpus for multi-document summarization.

3 Data Overview

We have compiled the Spotify Podcast Dataset, the first large scale corpus of podcast audio data with automatically generated transcripts. This corpus is drawn from a variety of creators, ranging from professional podcasters with high production value, to amateurs recording podcasts using an application on their mobile phone. The podcasts cover a wide range of topics including lifestyle & culture, storytelling, sports & recreation, news, health, documentary, and commentary. In addition, the content is delivered in a variety of structural formats, number of speakers, and levels of formality, some scripted, others improvised, and presented in the forms of narrative, conversation, or debate. Besides search and summarization, this data is valuable for tasks such as document segmentation or dialog modeling, and will enable new avenues of speech and language technology research.

Our corpus consists of over 100,000 podcast episodes, consisting of nearly 60,000 hours of audio and accompanying transcripts, as well as metadata such as creator-provided descriptions. The data was initially provided in the the context of the TREC Podcast Track (Jones et al., 2020). We now make it available for more general research use.

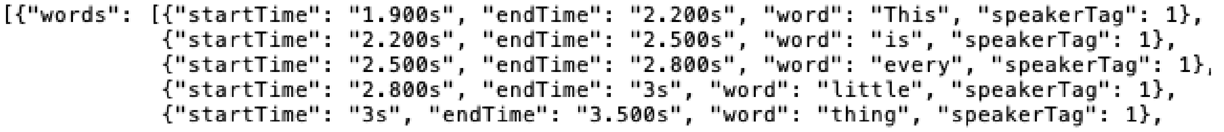
3.1 Data Sampling and Transcription

We randomly sampled 105,360 podcast episodes published between January 1, 2019 and March 1, 2020 from the Spotify platform. After filtering for several criteria shown in Table 1, we sampled about 10% from professional creators, with the remainder coming from amateur podcast creators. Podcast episodes were sampled uniformly at random. The episodes are all Spotify owned-and-operated, for copyright reasons. Currently the data set is restricted to the English language. We hope to extend the data set to further languages in the near future. The language determination is based on (1) the language indicated

[1] https://pypi.org/project/langid/

Language	We restricted our dataset to English according to the metadata tags provided by the creator. Since this labeling is somewhat noisy, we further filtered by running the n-gram based langid.py[1]
Length	We filter out any non-professionally published episodes that are longer than 90 minutes
Speech Presence	Using a proprietary speech detection algorithm, we ignored episodes belonging to podcasts that averaged less than 50% speech over the duration of the episode. This filters out podcasts that are more music than speech, or white noise and meditation.

Table 1: Filters used in sampling podcast episodes for the corpus.



```
[{"words": [{"startTime": "1.900s", "endTime": "2.200s", "word": "This", "speakerTag": 1},
            {"startTime": "2.200s", "endTime": "2.500s", "word": "is", "speakerTag": 1},
            {"startTime": "2.500s", "endTime": "2.800s", "word": "every", "speakerTag": 1},
            {"startTime": "2.800s", "endTime": "3s", "word": "little", "speakerTag": 1},
            {"startTime": "3s", "endTime": "3.500s", "word": "thing", "speakerTag": 1},
```

(a) Transcript snippet

Episode Name	Mini: Eau de Thrift Store
Episode Description	ELY gets to the bottom of a familiar aroma with cleaning expert Jolie Kerr. Guest: Jolie Kerr, of Ask a Clean Person. Thanks to listener Theresa.
Publisher	Gimlet
RSS Link	`https://feeds.megaphone.fm/elt-spot`

(b) Some of the accompanying metadata

Figure 1: Sample from an episode transcript and metadata

by the creator of the podcast as well as (2) a further automatic language identification algorithm applied to the creator-provided description. In spite of this we found a number of non-English podcasts in the dataset. This reflects how the multi-lingual reality of the data at hand defies the assumption of mono-lingual cultures: some descriptions given for non-English podcasts are written in English, from cultural areas where English frequently is more frequently used for writing; some other podcasts use many languages as a matter of course. Some examples of the types of multi-lingual podcasts episodes in the corpus include language learning podcasts, where English is the language of instruction, code-switching (eg Tagalog or Spanish speakers occasionally using English words and phrases), and podcasts analysis of religious texts where the text is read in the original language, and then the analysis of that text is in English. The podcast episodes cover a range of geographical regions, topical domains, and production quality; they vary in length and they include very short trailers as well as hour-long pieces.

We generate the text transcripts automatically using Google's Cloud Speech-to-Text API[2], which provides word-level time alignments for each word as well as speaker diarization, casing, and punctuation. Figure 1 shows an example snippet from a transcript and metadata, which includes episode name, show and episode description, publisher, duration, and the RSS header. The automatic speech recognition output showed robustness across the heterogeneous dataset, with a sample word error rate of 18.1% and a named entity recognition accuracy of 81.8%. This word error rate is higher than the output of highly optimized state of the art systems on corpora like Switchboard (Godfrey and Holliman, 1993) that report a word error rate of less than 5% (Bhatnagar et al., 2020), likely because of the domain mismatch between podcasts and the training data for the speech recogniser. However, we believe this word error rate is low enough that the transcribed corpus is valuable to the NLP, speech, and linguistics communities, as long as the noise is considered during algorithm development and analysis. Furthermore, since we do release the full audio as well, researchers that rely on clean transcripts may wish to manually transcribe the data. We also anticipate that the state of the art in automatic speech recognition will improve in the coming years, allowing for more accurate automatic transcriptions.

[2] `https://cloud.google.com/speech-to-text/docs/video-model`

3.2 Corpus Characteristics

The episodes in our corpus come from 18,376 podcast shows. 52% of shows are represented by more than one episode in the sample. The average episode duration is 33.8 minutes and 5,700 transcribed words, with large variance. Creator-provided episode descriptions average 85 words in length. The most to least common categories (as given by the creators in the RSS feed), weighted by episode length, are: Comedy, Sports, Health & Fitness, Society & Culture, and Education, Science, News & Politics, Government & Organization, and Fiction. The geographic origins of a small number (2,223) of these episodes are provided by the creators. Of those, majority (67%) come from the US, followed by Great Britain, Canada, Australia, and India.

Using the automatically inferred speaker diarization, the median speaker turn length per episode is about 110 seconds; more information on speaker distributions is in Appendix A. The automatic diarization is noisy: on manually checking 20 random episodes, we found that 11 have errors in the number of speakers, and another 4 have errors in speaker boundaries.

As an indication of the linguistic differences of the podcast data from traditional written corpora, a comparison with the Brown corpus (Francis and Kučera, 1967) shows how relative frequency of 1st person pronoun and amplifiers[3], features characteristic of conversational, informal language style, are much more common than in the Brown corpus (Table 2). This hints that this data may be of interest to research in sociolinguistics or computational social science.

Feature	Podcast data	Brown corpus
1st person pronouns	4.3%	0.40% (Press, reviews) - 2.6% (Romance novels)
Amplifiers	0.71%	0.15% (Press, reportage) - 0.35% (Press, reviews)

Table 2: Some selected lexical items' relative frequency of occurrence

Fitting an LDA topic model (Blei et al., 2003) with 100 topics to the transcripts yields topics corresponding to the categories and themes in the dataset, as well as discourse markers and slang reflecting the different styles (Table 3).

game play team ball point player win playing played three better line season ...	content
kid family mom child parent dad life old home mother house sister father ...	
god jesus church life lord word love bible christ heart spirit faith verse pray ...	
money pay dollar month business million property hundred paid real thousand ...	
song music album artist listen love record hip hop sound track new heard ...	
yeah oh okay yes exactly gonna feel guess sure cool pretty stuff definitely hmm ...	discourse
okay question yes maybe saying tell talk oh answer ask talking sure person thank point ...	
different example might use term important point change type level able may bit ...	

Table 3: A selection of LDA topics showing the breadth of both subjects (sports, family, religion, business, music, etc) and discourse styles (informal, conversational, technical, etc) in the dataset.

4 Search: Spoken Passage Retrieval

High-quality search of topical content of podcast episodes is challenging. Existing podcast search engines index the available metadata fields for the podcast as well as textual descriptions of the show and episode (Besser et al., 2008). These descriptions often fail to cover the salient aspects of the content. Improving and extending podcast search is limited by the availability of transcripts and the cost of automatic speech recognition. Our case-study is for fixed-length segment retrieval: given an arbitrary query (a phrase, sentence or set of words), retrieve topically relevant segments from the data. These segments can then be used as a basis for topical retrieval, for visualization, or other downstream purposes (Eskevich et al., 2012). A segment, for the purposes of our benchmark, is a two-minute chunk with one minute

[3] Amplifiers are a lexical items that increase the intensity of an expression, typically constructed as an adverbial, e.g. *very*, *really*, *totally*, or *amazing* (Quirk et al., 1985). The list used here is found in Appendix B.

Query	Type	Description
black hole image	topical	In May 2019 astronomers released the first-ever picture of a black hole. I would like to hear some conversations and educational discussion about the science of astronomy, black holes, and of the picture itself.
story about riding a bird	re-finding	I remember hearing a podcast that had a story about a kid riding some kind of bird. I want to find it again.
daniel ek interview	known item	Someone told me about a podcast interview with Daniel Ek, CEO of Spotify, about the founding and early days of Spotify. I would like to find the show and episode that contains that interview. Other interviews with Ek are relevant as well.

Table 4: Sample topics with query and description

overlap and starting on the minute; e.g. 0.0-119.9 seconds, 60.0-179.9 seconds, 120.0-239.9 seconds, etc. This creates 3.4M segments in total from the benchmark with the average word count of 340 ± 70.

4.1 Evaluation Data for Search

We created a small set of search information needs, called *topics*, following those used by the Text REtrieval Conference (TREC) (Voorhees and Harman, 2005). Each topic consists of a keyword query and a description of the user's information need. Topics can be one of three types: topical (general information about the topic), re-finding (searching for a specific episode the user heard before), and known item (finding something that is known to exist but under an unknown name) (Besser et al., 2010). Table 4 displays sample topics for each type.

Gold standard data for evaluation consists of human judgments of the relevance of segments to the topics. We used a simple BM25-based search to retrieve segments for judging, manually varying the query terms to try to increase coverage. We started with expert annotation by the paper authors on 609 passages retrieved for an initial set of 8 topics, then added 1060 crowd-sourced labels for passages retrieved for 14 more for a total of 22 topics, with annotations for 1669 query-passage pairs. To assist their judgment they could use the metadata, the full transcript, the audio, and any other resources they found helpful. The annotators used a standard graded scale of Excellent/Good/Fair/Bad, along with a Perfect grade for re-finding and known item topics. Table A1 in Appendix C shows the guidelines we provided the human assessors.

For collecting relevance judgements on the remaining 14 topics, we used the Appen[4] system for crowd-sourcing. We used our expert annotated judgements on the first 8 queries as the assessors' quality control tests for crowd-sourcing. We pooled the top 50 retrieved segments from the four aforementioned retrieval systems. Every segment was annotated by at least three annotators and in the case of disagreement we let the system to go up to 7 trusted annotations. These assessments proved to be quite noisy. To increase their utility, we only used judgments from assessors that had at least 40% accuracy in the quality control tests (i.e. 40% agreement with our own assessments, in line with Voorhees' work showing 40% agreement about relevance among expert assessors (Voorhees, 2000).

4.2 System Description for Search

We implemented as baselines standard retrieval models BM25 and query likelihood (QL) with the RM3 relevance model for relevance feedback (Lavrenko and Croft, 2017), using the Pyserini package[5] for search functionality, built on top of open-source Lucene[6] search library. Stemming was performed using the Porter stemmer. Four models, BM25, BM25+RM3, QL, and QL+RM3, are used with Anserini's default parameters.[7]

[4] `https://appen.com`

[5] `https://github.com/castorini/pyserini` – a Python front end to the Anserini open-source information retrieval toolkit (Yang et al., 2017)

[6] `https://lucene.apache.org`

[7] BM25 parameter settings $k = 0.9, b = 0.4$; RM3 settings $\mathit{fbTerms} = 10, \mathit{fbDocs} = 10, \mathit{originalQueryWeight} = 0.5$; QL setting for Dirichlet smoothing $\mu = 1000$

4.3 Results for Search

We use mean nDCG metric for evaluation in this task. An episode may contain one or more relevant segments, some of which may be overlapping, but these are treated as independent items for the purpose of nDCG computation. We evaluated each system over the 22 topics described above. Table 5 and Table 6 show results, with the former showing results broken out by query as well as overall mean, and the latter showing only the mean. Note that systems are not distinguishable; none of the results are statistically significant. However, we do consistently see that QL has the highest nDCGs, and both QL and BM25 have higher nDCGs than their RM3 counterparts.

		nDCG@5				nDCG@10			
		BM25	BM25+RM3	QL	QL+RM3	BM25	BM25+RM3	QL	QL+RM3
1	coronavirus spread	0.6655	0.6597	**0.7169**	0.5933	0.6717	**0.7278**	0.678	0.6579
2	greta thunberg cross atlantic	0.5801	0.1461	**0.8136**	0.4469	0.4742	0.2731	**0.5655**	0.391
3	black hole image	**0.8721**	0.851	0.7261	0.7104	**0.7921**	0.785	0.7325	0.7413
4	story about riding a bird	0	0	0	0	0	0	0	0
5	daniel ek interview	0	0	0	0	0	0	0	0
6	michelle obama becoming	**0.0838**	0	0	0	**0.0643**	0	0.0363	0
7	anna delvey	0	0	0	0	0	0	0	0
8	facebook stock prediction	0.5591	0.3367	**0.7016**	0.4409	0.6005	0.5394	**0.6792**	0.5477
	all	0.3451	0.2492	**0.3698**	0.2739	0.3253	0.2907	**0.3364**	0.2922

Table 5: nDCG scores for 8 human expert annotated topics.

	nDCG@5	nDCG@10
BM25	0.2737	0.3325
BM25+RM3	0.2731	0.3261
QL	0.2660	0.3357
QL+RM3	0.2542	0.3329

Table 6: nDCG scores for 14 crowdsourced test topics.

4.4 Lessons Learned for Spoken Passage Retrieval

Among the IR systems we tested, we do not observe significant difference in performance, likely due to the limitations of basic bag-of-word strategies. However, Table 5 shows different test topics achieve very different results. Three queries retrieve no relevant material; one retrieves very little. Two queries suffer from automatic speech recognition errors, as they create challenges for retrieving named entities. For example, we observed that *anna delvey* is never transcribed correctly, but similar-sounding phrases like *in a del v*, and *an adele v* are found in the transcripts instead. Similarly, *ek* is often mistranscribed as *ech* or *eck*. Systems will need to be more robust in retrieving low confidence named entities in the presence of automatic speech recognition errors.

The fourth query *story about riding a bird* is not well suited to traditional query-term matching information retrieval techniques. This suggests an approach involving classifying podcasts into types, eg story, interview etc, then recognizing the type sought by a query. The sixth query *michelle obama becoming* is hurt due to the common word *becoming* and the relatively high frequency with which Michelle Obama is a subject of discussion in podcast episodes. Advanced query-processing bringing in real-world knowledge that Michele Obama is the author of the book *Becoming* could address this case. We also find that the documents in languages other than English (Table 1) can become distractors: when run through automatic speech recognition for English they produce many less-frequent terms which can be retrieved despite being irrelevant to the query.

One interesting observation with our pseudo relevance expansion experiments is the "poison pill" effect of the expansion terms using RM3 (Terra and Warren, 2005). For almost all of our queries, exploiting RM3 for extracting expansion terms degraded the retrieval performance. Error analysis of query number 2 shows that terms related to *atlantic* (such as *shark*, etc.) are boosted whereas terms related to *greta thunberg* are lowered.

Length	descriptions that are very long (> 750 characters) or short (< 20 characters) amounting to $24,033$ or 23% of the descriptions.
Similarity to other descriptions	descriptions with high lexical overlap (over 50%) with other episode descriptions amounting $15,375$ or 15% of the descriptions.
Similarity to show description	descriptions with high lexical overlap (over 40%) with their show description, amounting to $9,444$ or 9% of the descriptions.

Table 7: Filters to remove less descriptive episode descriptions, to form the *brass subcorpus*.

5 Summarization

Automated document summarization is the task of condensing an input text into a much shorter form that preserves most of the salient information. This dataset presents several challenges: 1) the input documents are automatically transcribed, and thus subject to speech recognition errors, 2) the documents are frequently of a casual, conversational nature, with utterance fragments and disfluencies, and 3) the documents are significantly longer than typical summarization data. Thus, this task is most closely related to prior work in spoken document summarization and long document summarization (Cohan et al., 2018; Xiao and Carenini, 2019).

5.1 Data Preparation for Summarization: Brass Subcorpus and Gold Test Data

To train supervised models on this dataset, we consider the creator-generated descriptions as our reference summaries. However, these descriptions vary widely in quality and are not always intended to act as summaries of the episode content, reflecting the different uses creators have for descriptions and the different genres of podcast in the sample. In order to select a subset of the corpus that is suitable for training supervised models, we filtered the descriptions using three heuristics shown in Table 7. These filters overlap to some extent, and remove about a third of the entire set. The remaining 66,245 descriptions we call the *Brass Set*.

To derive gold labeled data, we internally annotated the outputs of different baseline systems on a sample of 303 episodes. We asked annotators to assess a summary's quality on a Excellent/Good/Fair/Bad (EGFB) scale, after reading the full transcript and/or listening to some of the audio if needed. Table A2 in Appendix C shows the guidelines we used.

5.2 Baseline Systems: Unsupervised Extractive and Supervised Abstractive

We ran an unsupervised summarizer, TextRank (Mihalcea and Tarau, 2004)[8], on the test data. The algorithm creates a graph of sentences, where the edge between a pair of sentences represents their similarity, and the sentences of highest importance, or "centrality", are computed using PageRank. We extract the top two central sentences as the unsupervised summary.[9] We also generated a naive baseline consisting of the first minute of spoken content.

We ran two variants of supervised models for generating abstractive summaries, both using BART (Lewis et al., 2020), as implemented in Huggingface[10]. For the first supervised variant, we simply used a pretrained model[11], which we refer to as BART-CNN, consisting of a large unsupervised BART model that was fine-tuned to the summarization task on the CNN/DailyMail dataset[12]. For our second supervised variant, we further fine-tuned the BART-CNN model to the podcast data, using the brass training set. We refer to this model as BART-PODCASTS. For both of these, we used the default hyperparameter settings, including a maximum input length requirement of 1024 tokens, significantly shorter than the average transcript length (thus, for longer inputs, the model simply ignored everything after the first 1024 tokens).

[8] We used the Python `sumy` package, `https://github.com/miso-belica/sumy`

[9] We also ran LexRank (Erkan and Radev, 2004) and a summarizer using LSA (Steinberger and Jezek, 2004), but found from a pilot evaluation that TextRank was more successful.

[10] `https://github.com/huggingface/transformers/tree/master/examples/summarization`

[11] `https://huggingface.co/facebook/bart-large-cnn`

[12] `https://s3.amazonaws.com/datasets.huggingface.co/summarization/cnn\_dm.tgz`

5.3 Evaluating Summary Quality

For evaluation of the baseline system outputs, we consider both automated metrics and human assessments. For automated metrics, we use standard flavors of ROUGE, as implemented in FILES2ROUGE[13] using the (noisy) creator descriptions as the reference.

Despite the variance in quality of the creator descriptions, we present the ROUGE scores against these descriptions as reference summaries and compare them against human judgements. We give the ROUGE scores on the test set broken out separately into the set of episodes whose descriptions passed the brass set filter versus those that failed the filter in Table 8.

	Brass			Non-Brass		
	R1-F	R2-F	RL-F	R1-F	R2-F	RL-F
FIRST MINUTE	18.90	3.92	9.68	16.89	3.67	9.78
TEXTRANK	15.25	2.04	8.69	13.04	1.58	7.99
BART-CNN	20.67	4.87	12.6	22.93	5.3	14.52
BART-PODCASTS	28.24	13.34	21.39	29.46	12.87	22.07

Table 8: ROUGE scores bucketed by whether the test descriptions passed the brass filter.

The ROUGE scores are in the same range as other reported experiments on this dataset (Zheng et al., 2020; Jones et al., 2020). They are lower than many other summarization benchmarks such as those on news corpora, for several likely reasons: (1) we do not have true reference summaries and the creator descriptions that we use as references were not written with the intent to summarize the podcast, (2) the transcripts are noisy, (3) the informality and heterogeneity of many podcasts makes them difficult to summarize.

To obtain assessments for the summary outputs, we asked human assessors to provide judgements assessed against the transcript, rather than against a gold summary. The results (Table 9) are robust: both the BART-CNN and BART-PODCASTS summarizers are nearly as good as the creator-provided descriptions on average, and in many specific cases provides better and more useful output. The unsupervised methods are rated lowest, with the FIRST MINUTE baseline outperforming TEXTRANK, likely since the first minute of podcasts often describes the content to follow.

	Brass					Non-Brass				
	E	G	F	B	Pct Good or Better	E	G	F	B	Pct Good or Better
CREATOR	37	35	33	39	50%	36	57	36	30	58%
FIRST MINUTE	10	33	46	55	30%	10	38	61	50	30%
TEXTRANK	2	10	43	89	8%	3	13	35	108	10%
BART-CNN	8	47	40	49	38%	28	40	41	50	43%
BART-PODCASTS	17	50	43	34	47%	37	51	35	36	55%

Table 9: Human labeled score distribution

5.4 Analysis of Summarization Results

In order to understand how well the brass labeled set will work as an automated training or test set, we analyze the quality with expert labels. We see from Table 9 that creator descriptions, taken as summaries, are of variable quality and that the summaries generated by supervised models have comparable performance. We also see that surprisingly, the nearly on-par performance of BART-PODCASTS holds for both the brass and the non-brass set. For more discussion and examples of this, see Appendix Section E.

The correlation between ROUGE and human judgements can degrade in spoken domains with multiple speakers (Liu and Liu, 2008). This issue could be further exacerbated in this podcast dataset, where our reference summaries are the noisy creators' episode descriptions. However, we find the same ranking of models by manual annotations and ROUGE scores: BART-PODCASTS $>$ BART-CNN $>$ FIRST MINUTE $>$ TEXTRANK. To test this further, we grouped the description by their human labels, and compared the

[13] https://github.com/pltrdy/files2rouge

induced system rankings of those with Excellent/Good descriptions as references to those with Fair/Bad reference descriptions. We found that the same ranking between systems holds across these buckets; for details, see Tables A5 and A6 in the Appendix. This suggests that ROUGE scores are meaningful for automated evaluation. We plan on further analysis using a larger human labeled set in the future.

On the whole, the abstractive BART models were rated higher than TextRank and the first-minute baseline on both human and ROUGE evaluations. Extractive models suffer from errors caused by speech recognition or the natural disfluency of spoken language, whereas the abstractive models seem to be more able to generalize over these errors and generate relatively fluent written language. Furthermore, while extractive models pick out topically salient bits of the transcript, those isolated bits do not always translate to an overview of the episode, whereas the abstractive models are able to generate overview statements from the transcript (example 1 in Table A7). Extractive models also suffer from failing to contextualize the text they select.

6 Conclusions and Future Work

We have presented the first large-scale dataset of transcribed podcasts. With this we have given benchmarks for a passage retrieval and a summarization task, along with an analysis that highlights ways in which this widely-varying spoken domain presents challenges for natural language processing and information retrieval.

In this work, we have limited our analysis to the transcriptions; however, there is much to be gained from considering the audio data as well for these and other tasks. In the NLP domain, podcasts are an ideal testbed not only for retrieval and summarization from transcripts, but also end to end summarization – translating the original audio into either a written summary or a short audio trailer – or retrieval tasks that leverage the audio, such as keyword search and spoken document retrieval. Given that there are multiple interlocutors in a podcast, speaker identification and role prediction are relevant problems of interest. In the information retrieval domain, the collection presents challenges in the noisy nature of the data, as well as the highly varied ways of speaking. The range of topics, stances, sentiments, and conversation styles that are present in the corpus provide rich ground for opinion mining and discourse analysis. Podcasts are also a promising medium for developing models that consider linguistic style in addition to topical material. The very varied styles and topics in the corpus suggest that this data may be of interest to research in sociolinguistics or computational social science. Paired with the audio files, they are also a resource for speech processing and the study of the acoustic aspects of the domain.

References

James Allan, Jaime Carbonell, George. Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic Detection and Tracking Pilot Study: Final Report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne. 007.

Jana Besser, Katja Hofmann, and Martha A Larson. 2008. An Exploratory Study of User Goals and Strategies in Podcast Search. In *Proceedings of Workshop-Woche: Lernen, Wissen & Adaptivität (LWA).*

Jana Besser, Martha Larson, and Katja Hofmann. 2010. Podcast search: User goals and retrieval technologies. *Online information review.*

Aadyot Bhatnagar, Caiming Xiong, Guangsen Wang, Richard Socher, Weiran Wang, and Yingbo Zhou. 2020. An investigation of phone-based subword units for end-to-end speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH).*

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3.

John W. Du Bois and Robert Engebretson. 2005. Santa Barbara corpus of spoken American English. *Linguistic Data Consortium.*

Alexandra Canavan, David Graff, and George Zipperlen. 1997. Callhome American English speech. *Linguistic Data Consortium.*

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans. ACL.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22.

Maria Eskevich, Walid Magdy, and Gareth JF Jones. 2012. New metrics for meaningful evaluation of informally structured speech retrieval. In *European Conference on Information Retrieval (ECIR)*. Springer.

Marcello Federico and Gareth JF Jones. 2003. The CLEF 2003 cross-language spoken document retrieval track. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 646–652. Springer.

W Nelson Francis and Henry Kučera. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, David S Pallett, Nancy L Dahlgren, and Victor Zue. 1990. TIMIT acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*.

John S Garofolo, Cedric GP Auzanne, and Ellen M Voorhees. 2000. The TREC Spoken Document Retrieval Track: A Success Story. *NIST SPECIAL PUBLICATION*, 500(246).

John S. Garofolo, Jonathan Fiscus, and Audrey Le. 2004a. Rich Transcription Broadcast News and Conversational Telephone Speech. *Linguistic Data Consortium*.

John S Garofolo, Christophe Laprun, Martial Michel, Vincent M Stanford, and Elham Tabassi. 2004b. The NIST meeting room pilot corpus. In *Proceedings of the 4th Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-Scale Multi-Document Summarization Dataset from the Wikipedia Current Events Portal. In *Proceedings of the 58th Meeting of the Association for Computational Linguistics (ACL)*. ACL.

John J. Godfrey and Edward Holliman. 1993. Switchboard-1 release 2. *Linguistic Data Consortium*.

Yoichiro Hasebe. 2015. Design and implementation of an online corpus of presentation transcripts of TED talks. *Procedia-Social and Behavioral Sciences*, 198.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop*, Hidden Valley, Pennsylvania.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI Meeting Corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*.

Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. 2020. Overview of the TREC 2020 Podcasts Track. In *The 29th Text Retrieval Conference (TREC 2020) Notebook*. NIST.

Konstantinos Koumpis and Steve Renals. 2005. Automatic summarization of voicemail messages using lexical and prosodic features. *ACM Transactions on Speech and Language Processing (TSLP)*, 2.

Victor Lavrenko and W. Bruce Croft. 2017. Relevance-based language models. *SIGIR Forum*, 51(2).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL.

Feifan Liu and Yang Liu. 2008. Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (ACL)*, Columbus, Ohio. ACL.

Iain Mccowan, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The AMI meeting corpus. In *Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*. Noldus.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL.

Derek Miller. 2019. Leveraging BERT for Extractive Text Summarization on Lectures. *arXiv preprint arXiv:1906.04165*.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of contemporary English*. London: Longman.

Damiano Spina, Johanne R. Trippas, Lawrence Cavedon, and Mark Sanderson. 2017. Extracting audio summaries to support effective spoken document search. *Journal of the Association for Information Science and Technology*, 68(9).

Josef Steinberger and Karel Jezek. 2004. Using Latent Semantic Analysis in Text Summarization and Summary Evaluation. In *Proceedings of International Conference on Information Systems Implementation and Modelling*.

Paul Tardy, David Janiszek, Yannick Estève, and Vincent Nguyen. 2020. Align then Summarize: Automatic Alignment Methods for Summarization Corpus Creation. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*. European Language Resources Association.

Egidio Terra and Robert Warren. 2005. Poison pills: harmful relevant documents in feedback. In *Proceedings of the 14th ACM international conference on Information and knowledge management CIKM*. ACM.

Ellen M Voorhees and Donna K Harman. 2005. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press, Cambridge.

Ellen M Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management*, 36.

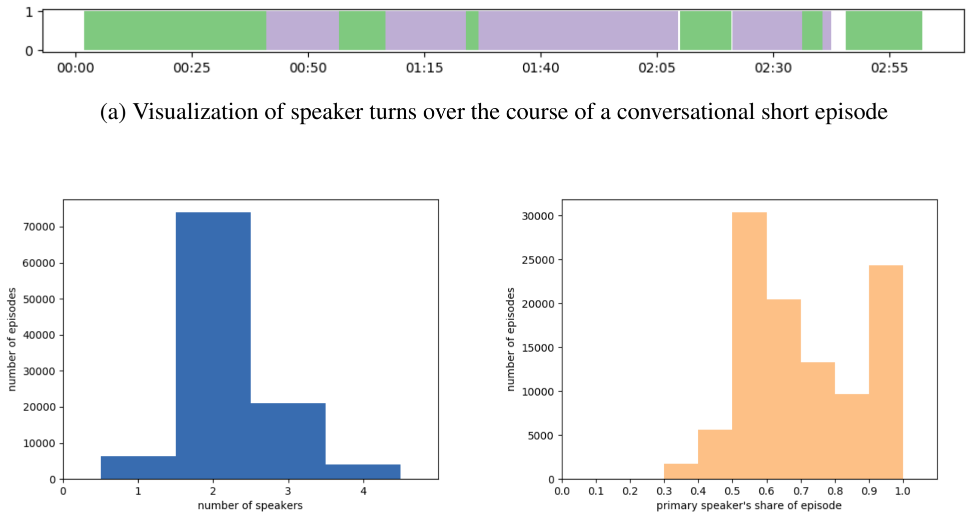
Gavin Whitner. 2020. The meteoric rise of podcasting. https://musicoomph.com/podcast-statistics (Accessed July 2020).

Wen Xiao and Giuseppe Carenini. 2019. Extractive Summarization of Long Documents by Combining Global and Local Context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong. ACL.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1253–1256.

Chujie Zheng, Harry Jiannan Wang, Kunpeng Zhang, and Ling Fan. 2020. A Baseline Analysis for Podcast Abstractive Summarization. *arXiv preprint arXiv:2008.10648*.

A Appendix: Speaker distributions



(a) Visualization of speaker turns over the course of a conversational short episode

(b) Number of speakers per episode.

(c) Primary speaker's share.

Figure A1: The dataset comprises episodes ranging from monologues to multi-speaker conversations. The plots are derived from the automatic speaker diarization output. While the output may be noisy, the aggregate distributions demonstrate the different conversational styles in the dataset.

B Appendix: List of amplifiers

absolutely
amazing
amazingly
awfully
completely
definitely
dramatic
dramatically
drastic
drastically
emphatic
emphatically
enormously
entirely
exceedingly
exceptional
exceptionally
excessively
extraordinarily
extraordinary
extremely
famously
fantastic
fantastically
genuinely
greatly
highly
horribly
hugely
immaculately
immensely
incredible
incredibly
insanely
intensely
overly
particularly
perfectly
phenomenal
phenomenally
radically
really
remarkable
remarkably
ridiculously
severely
significantly
striking
strikingly
strongly
substantially
surely
surprising
surprisingly
terribly
thoroughly
totally
truly
undoubtedly
unusual
unusually
utterly
vastly
very
wildly
wonderfully

C Appendix: Guidelines for assessment

*P*erfect	Should only be used for ”known item” and ”refinding” topic types with a specified ”Perfect” result. That result (and no other) should be judged ”Perfect”. For known item queries only: “perfect” for a point very near the start of the one relevant episode, and degrading from there if it’s in the episode but further away from the start, to fair if it’s the same show but not the right episode, to bad if it’s not even the same show.
*E*xcellent	The segment conveys highly relevant information, is an ideal entry point for a human listener, and is fully on topic. An example would be a segment that begins at or very close to the start of a discussion on the topic, immediately signalling relevance and context to the user.
*G*ood	The segment conveys highly-to-somewhat relevant information, is a good entry point for a human listener, and is fully to mostly on topic. An example would be a segment that is a few minutes “off” in terms of position, so that while it is relevant to the user’s information need, they might have preferred to start two minutes earlier or later.
*F*air	The segment conveys somewhat relevant information, but is a sub-par entry point for a human listener and may not be fully on topic. Examples would be segments that switch from non-relevant to relevant (so that the listener is not able to immediately understand the relevance of the segment), segments that start well into a discussion without providing enough context for understanding, etc.
*B*ad	The segment is not relevant.

Table A1: Guidelines for assessment of search relevance.

*E*xcellent	Accurately conveys all the most important attributes of the episode, which could include topical content, genre, and participants. It contains almost no redundant material which isn’t needed when deciding whether to listen.
*G*ood	Conveys most of the most important attributes and gives the reader a reasonable sense of what the episode contains. Does not need to be fully coherent or well edited. It contains little redundant material which isn’t needed when deciding whether to listen.
*F*air	Conveys some attributes of the content but gives the reader an imperfect or incomplete sense of what the episode contains. It may contain some redundant material which isn’t needed when deciding whether to listen.
*B*ad	Does not convey any of the most important content items of the episode or gives the reader an incorrect sense of what the episode contains. It may contain a lot of redundant information that isn’t needed when deciding whether to listen to the episode.

Table A2: Guidelines for assessment of summaries.

D Appendix: Full ROUGE scores

	R1-R	R1-P	R1-F	R2-R	R2-P	R2-F	RL-R	RL-P	RL-F
FIRST MINUTE	14.45	41.63	18.90	3.0	9.13	3.92	7.16	24.52	9.68
TEXTRANK	12.1	30.62	15.25	1.64	4.26	2.04	6.78	18.71	8.69
BART-CNN	26.4	22.7	20.67	6.58	5.51	4.87	15.79	14.8	12.6
BART-PODCASTS	39.42	28.59	28.24	18.06	14.08	13.34	29.09	22.38	21.39

Table A3: ROUGE scores for 144 test descriptions that passed the brass filter.

	R1-R	R1-P	R1-F	R2-R	R2-P	R2-F	RL-R	RL-P	RL-F
FIRST MINUTE	11.23	45.71	16.89	2.39	11.09	3.67	6.38	29.69	9.78
TEXTRANK	9.11	35.44	13.04	1.12	4.35	1.58	5.52	23.04	7.99
BART-CNN	23.35	29.2	22.93	5.3	7.13	5.3	14.39	19.57	14.52
BART-PODCASTS	34.73	31.35	29.46	15.04	13.73	12.87	25.67	24.02	22.07

Table A4: ROUGE scores for the 159 test descriptions that did not pass the brass filter

	R1-R	R1-P	R1-F	R2-R	R2-P	R2-F	RL-R	RL-P	RL-F
FIRST MINUTE	9.1	43.74	13.51	2.28	11.46	3.41	5.29	30.26	8.09
TEXTRANK	8.35	30.52	11.38	1.31	4.3	1.71	5.2	20.89	7.23
BART-CNN	17.93	26.72	18.29	4.19	6.17	4.17	11.13	18.6	11.74
BART-PODCASTS	31.69	34.59	28.58	17.9	18.93	15.9	25.96	29.0	23.6

Table A5: ROUGE scores against the test descriptions were assessed by humans as bad or fair.

	R1-R	R1-P	R1-F	R2-R	R2-P	R2-F	RL-R	RL-P	RL-F
FIRST MINUTE	15.84	43.86	21.51	3.04	9.14	4.15	8.0	24.76	11.14
TEXTRANK	12.4	35.4	16.4	1.44	4.36	1.9	6.92	21.07	9.27
BART-CNN	30.56	25.64	24.86	7.37	6.55	5.9	18.37	16.26	15.2
BART-PODCASTS	41.53	26.37	29.24	15.48	9.89	10.9	28.6	18.59	20.34

Table A6: ROUGE scores against the test descriptions were assessed by humans as excellent or good.

D.1 Do episodes with better descriptions have better summaries?

We see that the ROUGE scores of all systems tend to be higher on episodes with Excellent or Good descriptions (Table A6) compared to those with Fair or Bad descriptions (Table A5). This may be due to one of two reasons: a better description is more "summary-like", implying greater similarity to system-generated summaries, and episodes with good descriptions are also of higher production quality and fluency, resulting in better summarization performance.

E Appendix: Creator descriptions compared to the abstractive models

Table A7 contrasts some creator and system summaries. In example 1, the creator summary is well punctuated with correct casing, is syntactically accurate, and describes briefly the intent of the podcast and where it would be made available. In contrast, the fine-tuned BART generated summary has incorrect casing, lacks punctuation at sentence boundaries, and does not have the same comprehensional quality as the CREATOR summary.

However, in example 2, it is evident that lack of details in the creator summary makes it uninformative to a user. On the other hand, the fine-tuned BART summary syntactically and semantically conveys the content of the podcast succinctly, but misses out on details that could better introduce the podcast show to a casual reader. Similarly, example 3 shows a Bad creator summary where the text is elongated and details are insufficient with no hint of topic of discussion. On the other hand, the model summary introduces the hosts by their names and then gives additional context for the episode, followed by additional topics discussed in the episode.

1	CREATOR (Good)	Motivated by a lack of basic humility in the mainstream media, I will be employing the Podcast format and Instagram to deliver information about the recent outbreak of Coronavirus (COVID-19) to help you discern fact from fiction on your own to empower an informed personal response.
	TEXTRANK (Bad)	I think the number one thing right now is just the public has been lulled into a sense of complacency regarding, you know, the reality of biology, you know, and we've seen the the explosion of last 15 years of the mobile device, you know, like the iPhone or the Android and social media and I think that you know, the Google Facebook Apple iPhone so forth, you know model of the internet largely is to blame for the hiccup in Wall Street last week without that sort of Insanely rapid expansion and the way that information is distributed.
	BART-PODCASTS (Excellent)	In this episode, Dr. William Cairo-Bowl introduces the podcast and gives his background and background background on the basic science of the coronavirus. He also provides a brief overview of what to expect in future episodes.
2	CREATOR (Excellent)	Get comfortable, and drift off as you listen to Genesis 37, 39-40. Listen to Amber Lawton as she reads you the Bible at bedtime. Pleasant dreams.
	BART-PODCASTS (Fair)	Let me Soothe you 2 Sleep as we listen 2, by request, more from an interesting newly released version "Beginnings" by editors Elmer Towns & Lee Fredrickson. This is part 1 of 2. — Join us in Bible.
3	CREATOR (Bad)	Importance level : High
	BART-PODCASTS (Good)	Import of crude oil from Iran has been banned under the US sanctions imposed in November 2018. India has been allowed to continue importing crude oil at a reduced rate of up to 6 months but now the exemption has been extended to till the end of May 23, 2019. The impact of this decision on Indian Crude oil.
4	CREATOR (Bad)	JOSHHHHHHHHHHHHHHHHHHHHHH GORDOOOOOOOOOOOONNNNNNNNN
	BART-PODCASTS (Excellent)	Brycen and Alfie are back with a recap of the Patriots/Titans preseason game, as well as a preview of the Week 2 preseason game against the Texans. They also talk about the return of Josh Gordon and what to expect from him this season.

Table A7: Extractive systems like TextRank fail to generate summaries that are fluent and span the themes in the episode in comparison to the abstractive model. While creator summaries tend to be good on the whole, there are several cases where the abstractive model generates summaries that are rated higher.