The Spotify Podcasts Dataset

Ann Clifton
aclifton@spotify.com

Aasish Pappu
aasishp@spotify.com

Sravana Reddy
sreddy@spotify.com

Yongze Yu
yongzey@spotify.com

Jussi Karlgren
jkarlgren@spotify.com

Ben Carterette
benjaminc@spotify.com

Rosie Jones
rjones@spotify.com

Abstract

Podcasts are a relatively new form of audio media. Episodes appear on a regular cadence, and come in many different formats and levels of formality. They can be formal news journalism or conversational chat; fiction or non-fiction. They are rapidly growing in popularity and yet have been relatively little studied. As an audio format, podcasts are more varied in style and production types than, say, broadcast news, and contain many more genres than typically studied in video research. The medium is therefore a rich domain with many research avenues for the IR and NLP communities. We present the Spotify Podcasts Dataset, a set of approximately 100K podcast episodes comprised of raw audio files along with accompanying ASR transcripts. This represents over 47,000 hours of transcribed audio, and is an order of magnitude larger than previous speech-to-text corpora.

1 Introduction

Podcasts are a relatively new form of audio media. Episodes appear on a regular cadence, and come in many different formats and levels of formality. They can be formal news journalism or conversational chat; fiction or non-fiction. They are sharply growing in popularity and yet have been relatively little studied. As an audio format, podcasts are more varied in style and production types than, say, broadcast news, and contain many more genres than typically studied in video research. The medium is therefore a rich domain with many research avenues for the Information Retrieval and Natural Language Processing communities.

In this work, we present the Spotify Podcasts Dataset, the first large scale corpus of podcast audio data with full transcripts. This corpus is drawn from a variety of heterogeneous creators, ranging from professional podcasters with high produc-tion values to amateurs without access to state-of-the-art production resources. The podcasts cover a wide range of topics including lifestyle and culture, storytelling, sports and 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, whether scripted or improvised, or in the form of narrative, conversation, or debate. These data present an interesting challenge for existing tasks such as spoken document retrieval, segmentation, and summarization, and will enable new avenues of speech and NLP research.

We release these data in collaboration with the TREC Text Retrieval Conference, along with a summarization and an ad-hoc retrieval challenge. For more information about participating in the challenge and obtaining the dataset , please see the TREC website[1].

2 Speech Datasets

Existing collections of spoken material and speech-to-text corpora have characteristics that differ importantly from this dataset. As noted above, podcasts extend spoken language communication in certain ways and collections of previous material do not capture that entire range of variation.

Existing collections have been collected for a variety of specific reasons, and for very specific domains, intended to capture regularities of some particular communication situation, such as e.g. the ATIS corpus of air travel information requests [Hemphill et al., 1990], meeting recordings [Garofolo et al., 2004b], telephone conversations [Canavan et al., 1997] and broadcast news [Garofolo et al., 2004a].

---

[1]https://trec.nist.gov/pubs/call2020.html

Further, many existing corpora present relatively clean audio scenarios, and from a limited selection of domains, often with a single speaker reading from a prepared text, such as the original TIMIT collection [Garofolo et al., 1990] or the many available 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]. These more formal settings or samples of formal content are useful for the study of acoustic qualities of human speech but represent a more idealised scenario than many practical audio processing tasks of interest today.

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 Englebretson, 2005] or 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.

This dataset represents the first large scale corpus of fully transcribed podcast data that we are aware of. Other transcribed podcast datasets of interest have only partial transcripts such as the one collated by Yang et al. [2019], who released a dataset of up to the first ten minutes of ~88K podcast episodes, with transcripts for ~46K of these, or are orders of magnitude smaller such as the Podcast Summaries data set of some 3,000 podcast episodes [Spina et al., 2017].

3 Data

This corpus consists of 100,000 podcast episodes, consisting of 50,000 hours of audio and accompanying transcripts. These episodes are randomly sampled across a wide range of regions, domains, and audio quality. The episodes vary in length and include some very short trailers to advertise other content. Table 1 gives descriptive statistics on word counts of podcast episodes in the dataset; Figure 1 gives the distribution over the lengths of the episodes in the dataset.

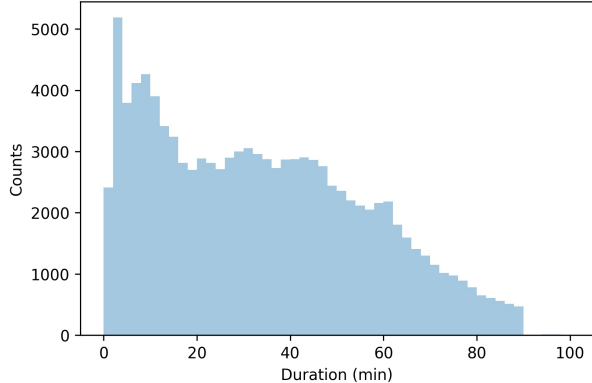|	min	average	max
minutes	<1	31.6	305.0
words	11	5,728	43,504

Table 1: Descriptive statistics

Figure 1: Distribution of episode duration. We eliminated most podcasts longer than 90 minutes from the collection. Out of the 100K+ episodes, there were 125 that exceeded 90 minutes in length (not shown in graph).

For each episode, we also supply the following metadata: show URI, episode URI, show name, episode name, show description, episode description, publisher, language, RSS link, and duration. In addition, we provide the full RSS header for each show, from which additional metadata can be extracted if desired.

3.1 Episode Selection

To select the episodes for inclusion in this dataset, we randomly sampled episodes published between January 1, 2019 and March 1, 2020, after filtering for several criteria.

For this dataset, we restricted episodes by language. While in future versions of this dataset we would like to provide multilingual data, in the current version we limit it to English. In order to do this, we first restricted the set of episodes using metadata tags as provided by the creator specifying the language of the show to be English. However, these tags can be inaccurate, so as an additional step, we run language identification[2] on the podcast description. We exclude the podcast title from the language identification step, as we found that titles are often very brief or named entities that yield low confidence predictions, whereas the descriptions are more frequently longer, grammatical

---

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

utterances. Since we recognize that many podcasts will have multilingual content and we wish to include this, we prefer to err on the side of precision when identifying non-English content, and therefore we only exclude those episodes that the identification tool positively identifies as being non-English.

We also did some filtering based on episode length. For the professionally produced podcasts (those from Spotify Studios, Parcast, and Gimlet), we did not limit by length, since these were distributed towards shorter lengths, with only 125 episodes exceeding 90 minutes and a maximum length of 305 minutes. However, for the non-professionally published episodes (those from Anchor), we restrict these in length to 90 minutes, as a heuristic quality filter, since excessively long amateur podcasts were more frequently subject to noise.

We further limited the dataset to include only those shows that had been streamed above a certain threshold number of times in the first month since their release. This heuristic provided us with a naive yet effective quality filter, removing defective or noisy material.

We also removed episodes that belonged to shows that averaged less than 50% speech overall. To do this, we used a classifier that automatically detects whether each one-second snippet of audio in the episode is noise, music, or speech.

Once these conditions were fulfilled, we then randomly sampled 100,000 episodes from the entire set.

3.2 Transcription

We generate the text transcripts automatically using Google Cloud Platform's Cloud Speech-to-Text API[3] (GCP-ASR), which provides word-level timestamps for each word. In addition, we include other inferred information it supplies: speaker diarization, casing, and punctuation insertion.

On a set of 1,600 English episodes, we conducted manual evaluation of the quality of the transcripts. We selected the episodes for the evaluation set to include a variety of domains, number of speakers, regional accents, level of formality, and audio quality. On our heterogeneous test set, GCP-ASR showed robustness across the dif-

---

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

ferences in the audio data quality and regional accents, with a word error rate of 18.1% and an accuracy for named entity recognition of 81.8%. Figure 2 shows an example snippet from a transcript.

3.3 RSS Headers

For each podcast episode, we provide the publicly available show-level RSS header HTML file snapshot at the time of assembling these data. These headers contain a wide variety of show-level and episode-level metadata, including the channel, title, description, author, link, copyright, language, image. We advise that these RSS headers change over time, and are subject to noise. For example, RSS headers may contain inaccurate episode-level information, such as links which are outdated. Thus, while the transcripts we provide match the audio files in this dataset, they may not match pointers in the RSS feed.

Figure 3 shows an example of a podcast episode's metadata from an excerpt from its RSS feed, which includes a pointer to the audio.

4 Conclusion

We have presented the Spotify Podcasts Dataset, the first large-scale dataset of podcasts with transcripts. We hope that this will encourage researchers to explore the podcast medium in speech and text processing tasks, such as information retrieval and summarization.

References

John W. Du Bois and Robert Englebretson. Santa barbara corpus of spoken american english. *Linguistic Data Consortium*, 2005. URL https://catalog.ldc.upenn.edu/LDC2005S25.

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

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

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, David S Pallett, Nancy L Dahlgren, and Victor Zue. Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*, 1990. URL https://catalog.ldc.upenn.edu/LDC93S1.

John S Garofolo, Cedric GP Auzanne, and Ellen M Voorhees. The trec spoken document retrieval track:

```
[{"words": [{"startTime": "0.900s", "endTime": "1.400s", "word": "Welcome", "speakerTag":
1}, {"startTime": "1.400s", "endTime": "1.500s", "word": "to", "speakerTag": 1},
{"startTime": "1.500s", "endTime": "1.700s", "word": "the", "speakerTag": 1},
{"startTime": "1.700s", "endTime": "2.100s", "word": "AIA", "speakerTag": 1},
{"startTime": "2.100s", "endTime": "2.800s", "word": "Vitality", "speakerTag": 1},
{"startTime": "2.800s", "endTime": "3.100s", "word": "One", "speakerTag": 1},
{"startTime": "3.100s", "endTime": "3.400s", "word": "Minute", "speakerTag": 1},
{"startTime": "3.400s", "endTime": "4.200s", "word": "Podcast.", "speakerTag": 1},
```

Figure 2: transcript snippet

```
<channel>
        <title><![CDATA[Grow Well]]></title>
        <description><![CDATA[A houseplant meets personal growth podcast by Dusty Hegge (that's me!). There are such beautiful parallels between
plant growth and human growth and I KNOW that is by God's design. So, on Grow Well, we talk about both! Everything from soulful personal growth to how to
keep your fiddle leaf fig alive - we'll talk about it all!]]></description>
        <link>http://dustyhegge.com/grow-well</link>
        <image>
                <url>https://d3t3ozftmdmh3i.cloudfront.net/production/podcast_uploaded_nologo/1368334/1368334-1580847055377-
f7cc298ca24c3.jpg</url>
                <title>Grow Well</title>
                <link>http://dustyhegge.com/grow-well</link>
        </image>
        <generator>Anchor Podcasts</generator>
        <lastBuildDate>Wed, 25 Mar 2020 21:56:43 GMT</lastBuildDate>
        <atom:link href="https://anchor.fm/s/8c07ff8/podcast/rss" rel="self" type="application/rss+xml"/>
        <author><![CDATA[Dusty Hegge]]></author>
        <copyright><![CDATA[Dusty Hegge]]></copyright>
        <language><![CDATA[en-us]]></language>
        <atom:link rel="hub" href="https://pubsubhubbub.appspot.com/"/>
        <anchor:support>https://anchor.fm/grow-well/support</anchor:support>
        <anchor:station>8c07ff8</anchor:station>
```

Figure 3: RSS header snippet

A success story. *NIST SPECIAL PUBLICATION*, 500(246):107–130, 2000.

John S. Garofolo, Jonathan Fiscus, and Audrey Le. Rich transcription broadcast news and conversational telephone speech . web download. *Linguistic Data Consortium*, 2004a. URL https://catalog.ldc.upenn.edu/LDC2004S11.

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

Yoichiro Hasebe. Design and implementation of an online corpus of presentation transcripts of ted talks. *Procedia-Social and Behavioral Sciences*, 198:174–182, 2015.

Charles T Hemphill, John J Godfrey, and George R Doddington. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

Damiano Spina, Johanne R. Trippas, Lawrence Cavedon, and Mark Sanderson. Extracting audio summaries to support effective spoken document search. *Journal of the Association for Information Science and Technology*, 68(9):2101–2115, 2017. ISSN 2330-1643. doi: 10.1002/asi.23831. URL http://dx.doi.org/10.1002/asi.23831.

Longqi Yang, Yu Wang, Drew Dunne, Michael Sobolev, Mor Naaman, and Deborah Estrin. More than just words: Modeling non-textual characteristics of podcasts. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 2019.