Leveraging multimodal content for podcast summarization

Leveraging multimodal content for podcast summarization

Lorenzo Vaiani, Moreno La Quatra, Luca Cagliero and Paolo Garza
Politecnico di Torino
Turin, Italy
{lorenzo.vaiani,moreno.laquatra,luca.cagliero,paolo.garza}@polito.it

ABSTRACT

Podcasts are becoming an increasingly popular way to share streaming audio content. Podcast summarization aims at improving the accessibility of podcast content by automatically generating a concise summary consisting of text/audio extracts. Existing approaches either extract short audio snippets by means of speech summarization techniques or produce abstractive summaries of the speech transcription disregarding the podcast audio. To leverage the multimodal information hidden in podcast episodes we propose an end-to-end architecture for extractive summarization that encodes both acoustic and textual contents. It learns how to attend relevant multimodal features using an ad hoc, deep feature fusion network.

The experimental results achieved on a real benchmark dataset show the benefits of integrating audio encodings into the extractive summarization process. The quality of the generated summaries is superior to those achieved by existing extractive methods.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Machine learning**; • **Information systems** → **Multimedia information systems**;

KEYWORDS

Podcast summarization, multimodal learning, extractive summarization, deep learning, multimodal features fusion

1 INTRODUCTION

Podcasts are episodic series of spoken-word digital audio files that users can easily download on their own devices and listen to. Their ever-increasing diffusion has recently stimulated the attention of the research community [15]. To improve the fruition of podcast content an appealing research branch has been devoted to podcast summarization. It entails extracting a summary consisting of a shortlist of text/audio snippets that are representative of the whole podcast content [13].

The need for summarizing podcasts is mainly related to the high variability in duration and content. Specifically, podcast duration ranges from a few minutes to even one hour or more (see, for example, the audio statistics on the Spotify dataset [5] available in Table 2). Furthermore, the topics covered by the podcasts significantly vary across different episodes and shows. To quickly access the key podcast information listeners could be interested in listening to a short trailer consisting of an extractive audio/text summary. Finally, the recurrent presence of advertisements throughout the podcasts fosters the use of summarization techniques to quickly access the key information in the podcast.

In a nutshell, podcast summarization is appealing for the following reasons: (1) Podcasts are extremely variable in topic and production style. (2) Podcast episodes are rather heterogeneous (see, for example, the statistics reported in Table 1). (3) Podcast content is often noisy due to the presence of overlapping speech, background effects, and advertisements [23].

Preliminary attempts to address podcast summarization have already been made. They produce two different kinds of outputs: (a) a textual summary of the audio speech transcription generated by using abstractive neural summarization methods (e.g.,[15, 25, 28]), or (b) an audio summaries extracted using speech summarization techniques (e.g.,[11, 31, 34]). A detailed literature review is given in Section 2.

Textual and audio sources are likely to provide complementary information. For example, the textual content misses the tone of voice of the speakers and the presence of potentially relevant background effects. However, to the best of our knowledge, none of the existing podcast summarization techniques jointly exploits textual and acoustic information in the summarization process. Another remarkable issue is related to the use of abstractive summaries in textual form, which cannot be directly linked to the corresponding audio snippet. The aforesaid issues call for new multimodal, extractive podcast summarization methods.

We present **MATeR** (namely, **M**ultimodal **A**udio-**Te**xt **R**egressor), a novel multimodal podcast summarization approach. It encodes both acoustic and textual podcast contents by using state-of-the-art attention-based sequence embedding models [1, 24]. A deep network is trained to estimate the semantic similarity between the sentence-level textual content and the podcast description. Thanks to end-to-end network training both textual and acoustic encodings contribute to improve the quality of the textual summary. Finally, the extracted sentences are mapped to the corresponding audio snippets (see Section 4 for a more thorough description of the presented method). Thus, our approach also returns the audio snippets corresponding to the summary.

The experiments run on a benchmark dataset [5] have shown (i) substantial performance improvements of the multimodal model against text-only strategies and (ii) a statistically significant performance improvements against state-of-the-art extractive summarization methods. The results will be thoroughly described in Section 6.

2 RELATED WORK

Multimodal summarization. In recent years, several research studies have addressed the problem of multimodal data summarization. Some of them focus on summarizing synchronous multimedia contents such as the audio, text, and video sequences of a video-recorded meetings [8], the pictorial story-lines consisting of image-text pairs [33], and the social events described by the photos and videos posted by the users [3, 27]. Some other address the summarization of asynchronous/non-aligned multimedia contents

(e.g., [16, 35]). The present work addresses the multimodal podcast summarization task, which is a specific case of synchronous content summarization.

Multimodal summarization techniques can be further categorized according to the type of generated summaries. Specifically, extractive methods (e.g., [4, 16]) shortlist portions of existing content whereas abstractive ones generate ad hoc summary content (e.g., by applying data fusion [9] or attention-based architectures [35, 39]). Similar to [18], it investigates the joint use of textual and acoustic latent features for summarization purposes. Unlike [18], MATeR focuses on a specific type of multimodal source, i.e., the podcast, rather than on a set of spoken documents (i.e., a broadcast news corpus).

Podcast summarization. Podcast datasets usually comprise a mix of audio and textual data (see, for example, [5]). However, all the existing approaches to podcast summarization (e.g., [21, 25, 28, 37]) mainly rely on text summarization algorithms applied to the podcast's speech transcriptions thus ignoring potentially relevant acoustic features for podcast summarization. The task of summarizing podcasts has been also investigated in one of the tracks of the Text REtrieval Conference (TREC) [14]. Specifically, they apply neural abstractive models, which produce new sentences on top of the original audio transcription. For example, in [28] the authors apply transformer-based methods to synthesize a truncated transcript version. To overcome the limitations of BERT-based sentence encoders [6] while coping long text spans, [21] proposes a model that is specifically designed for dealing with long-span dependencies. Furthermore, it also studies the influences of various sentence filtering methods that precede the abstractive summarization stage. The approach presented by [25] incorporates genre information and named entities into an abstractive, supervised summarization process. The key idea is to tailor the summary to the actual podcast style in order to effectively manage podcast heterogeneity. Unlike [15, 25, 28], the approach presented in this paper exclusively relies on extractive techniques. The idea behind it is to jointly attend latent textual and audio features in the selection of the relevant textual content and then extract the audio segments corresponding to the shortlisted text spans. Therefore, unlike all abstractive methods the proposed approach inherently generates multimodal summaries.

Speech summarization. It entails processing the audio files directly to produce a summary consisting of a selection of audio snippets. Traditional approaches rely on feature classification (e.g.,[11]), semi-supervised learning (e.g.,[34]), and graph clustering. More recently, in [31] the authors leverage sequence-to-sequence models to attend relevant acoustic features for speech summarization. Although speech summarizers are potentially applicable to the podcast audio speech, the scope of the present work is rather different, i.e., we summarize a multimodal source consisting of both podcast audio and text.

3 PROBLEM STATEMENT

Given a set of N podcast episodes $\mathbf{P}=\{P_j\}$, the present work aims at generating an extractive summary $\mathcal{S}_{P_j}$ separately for each episode P_j for which the (human-generated) description is unknown. Summary generation is based on a model trained on the episodes for which the relative description is known. As discussed below, the summary consists of a shortlist of text-audio pairs in P_j.

The summarization method considers multimodal information consisting of audio and text snippets extracted from the podcast. Specifically, for each episode P_j we consider both the audio sequence a_j and the speech transcription t_j. A more detailed description of the real-world dataset considered in our study is reported in Section 5.

To enable sentence-level text summarization each speech transcription t_j is split into a set of disjoint sentences $\{s_i^j\}$. Since acoustic and textual sources are synchronized, we can map each sentence s_i^j to the corresponding audio snippet a_i^j. Hereafter, we denote as $\mathbf{TA}_j$ the set of text-audio pairs $\langle s_i^j, a_i^j \rangle$ used as reference multimodal content units for P_j. Moreover, let $\mathbf{TA} = \bigcup_{j=1}^{N} \mathbf{TA}_j$ be the set of text-audio pairs occurring in the episodes in $\mathbf{P}$.

Let f: $\mathbf{TA} \rightarrow r_i^j$ be a function, belonging to a function space $\mathcal{F}$, that maps each text-audio pair $\langle s_i^j, a_i^j \rangle$ in $\mathbf{TA}$ to a relevance score r_i^j. The relevance of a text-audio pair is estimated as the textual similarity between s_i^j and the (human-generated) podcast description of the corresponding podcast d_j, i.e.,

$$r_i^j = sim(s_i^j, d_j)$$

It quantifies the extent to which a given text-audio pair in P_j is worth being included in the P_j's summary. Hereafter, we will use the podcast description as reference annotation for supervised learning.

We define $\mathbf{TA}_{train} \subset \mathbf{TA}$ as the set of pairs of the podcast episodes for which the descriptions are known, whereas $\mathbf{TA}_{unlabeled} \subset \mathbf{TA}$ is the set of pairs of the podcast episodes without descriptions. We aim at estimating the relevance score of an arbitrary pair $\langle s_i^q, a_i^q \rangle \in \mathbf{TA}_{unlabeled}$ for which the corresponding description is unknown based on a predictive model built on the annotated pairs, i.e., the text-audio pairs in $\mathbf{TA}_{train}$. To this end, we define a regressor $\mathcal{R}$ that explores the function space $\mathcal{F}$ to address the following optimization task.

$$\mathcal{R} = \arg\min_{\mathrm{f} \in \mathcal{F}} \sum_{\langle s_i^j, a_i^j \rangle \in \mathbf{TA}_{train}} E\Big(r_i^j, f(\langle s_i^j, a_i^j \rangle)\Big)$$

The classical regression task has the goal of finding the function that minimizes the prediction loss $E(\cdot)$[1] over the training data.

The regressor outputs are conveniently used to address the podcast summarization task. Specifically, given an episode P_q, for which the description is unknown, and the relevance score estimates r_i^q associated with each text-audio pair in P_q, the summarizer returns the top-K pairs in P_q in order of decreasing relevance (where K is a user-specified parameter).

As discussed in Section 5, since the description length is approximately equal to the average sentence length hereafter we will tailor the summarization task to the retrieval of the top ranked sentence (i.e., K=1).

[1] Whenever not otherwise specified, hereafter we will use the Root Mean Square Error as reference loss function.

4 METHOD

We present **MATeR** (namely, **M**ultimodal **A**udio-**Te**xt **R**egressor), a novel Deep Learning architecture for multimodal podcast summarization.

The key property of MATeR is to consider both acoustic and textual features in the selection of the most representative text-audio snippets. MATeR performs end-to-end training of the neural network-based regressor formalized in Section 3. Specifically, it combines text and audio encodings in a multimodal fusion network that first estimates the sentence similarity score and then back-propagates the prediction error through the whole network layers to get accurate sentence relevance estimates.

A sketch of the MATeR architecture is depicted in Figure 1. It is designed to perform a simultaneous analysis of audio and text modalities in order to create effective and accurate representations of the input data. To this aim, two single-modal heads are jointly trained to build the audio and text encoded representations (see Sections 4.1 and 4.2, respectively). The generated text and audio encodings are then combined together and processed by a multimodal feature fusion network, which consists of a stack of fully connected layers. The network predicts the relevance of each text-audio snippet to the output summary (see Section 4.3).

4.1 Text encoding

Transformer architectures [32] are established attention-based models to generate contextualized sentence embeddings. Among them, BERT [6] is the most renowned sentence encoder. The performances of BERT-like architectures are top-level on sentence similarity.

For this reason, MATeR adopts BERT to encode the input text snippets [24] into 768-dimensional vectors using a mean pooling token-aggregation strategy.

4.2 Audio encoding

To encode audio samples into a fixed-size vector representation we rely on a recently proposed speech encoder, namely Wav2Vec 2.0 [1]. It integrates both convolutional and transformer layers to generate a contextual representation, which is trained using self-supervision. While the original architecture has been fine-tuned to perform *automatic speech recognition*, its encoded representations has been proven to be effective for a wide range of audio-related tasks [10]. Specifically, we fine-tune the Wav2Vec2 model for audio features extraction.

In compliance with text-based encodings [24], the audio encoder aggregates the contextual speech representation using mean-pooling to generate 768-dimensional vectors for variable-length audio samples. Among the available aggregation methods tailored to acoustic features, mean-pooling turned out to be among the best performers for speaker recognition [30].

4.3 Multimodal fusion network

We combine text and audio encodings into a multimodal, unified latent representation.

The fusion network takes as input the single-model heads' output, consisting of separate 768-dimensional dense vectors for audio and text.

The network consists of a stack of fully-connected (FC) layers with ReLU activation function leveraging concatenated representation to estimate description-driven relevance of the input pairs. In our experiments, we set the width of each fully connected layer and depth of the fusion network to 1536 and 3, respectively. The prediction loss is estimated using the Mean Square Error and back-propagated through the deep network.

5 DATASET OVERVIEW

The Spotify podcast dataset [5] consists of approximately 100,000 podcast episodes belonging to about 18,000 different shows. For each podcast it includes the audio file, the transcript of the speech, and some related metadata (e.g., the podcast's creator and human-generated description). We deem both speech transcripts and audio files relevant to summarize podcast episodes, as they potentially convey complementary information. For example, the background effects, the speaker characteristics (e.g., tone of voice), and the presence of advertisements can be detected from the audio files, whereas they are partly missing or hard to detect in the speech transcription.

The multimodal data contents available at the Spotify dataset [5] are already partitioned in a set of text-audio snippets. Throughout the experiments, we have considered the predefined text-audio content splits, albeit MATeR supports different sentence- or paragraph-level tokenizations as well.

For evaluation purposes we apply a hold-out validation on a representative data sample consisting of 10% of the input data as training set, 1% of data as validation set, and 1% as test set. Data samples are stratified over the podcast shows to avoid partitioning the episodes of the same show in the training/validation/test sets.

To allow experiment reproducibility, the code and data samples are publicly available in the MATeR project repository[2].

5.1 Description filtering

Textual descriptions are manually written by the podcast creators and often include one or more advertising phrases. These phrases are intended to sponsor the products/tools that supported podcast realization or to mention the social network profiles/websites of the podcast author/episode guests.

To avoid introducing a bias in summary content selection and evaluation, we apply a semi-automatic approach to remove advertisements, commercial and promotion contents from the descriptions prior to training the models.

Specifically, similar to [26], we first split the description content into shorter text snippets using the sentence tokenization provided by spaCy library [12]. Next, we train a classifier on a small subset of manually annotated descriptions to automatically recognize advertising content. To this end, description phrases are labeled as *advertisement* if they include advertisements/promotions, *other* otherwise. Advertising content is early pruned from the descriptions before running the summarization process. To automatically label the non-annotated content we fine-tune the BERT transformer for binary text classification [6].

[2]The MATeR project repository, including data, annotations, and empirical results, is available at https://github.com/MorenoLaQuatra/MATeR

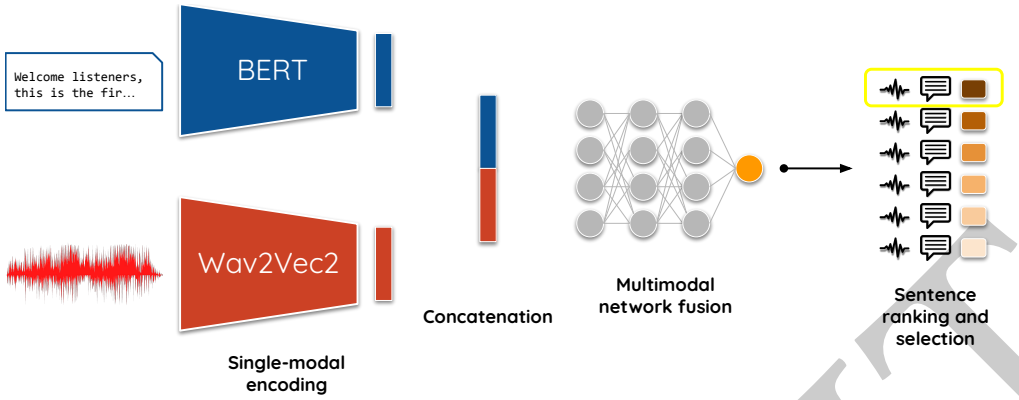


Figure 1: Sketch of the MATeR architecture.

Figure 2: Sentence scoring step.

To make the experiments fully reproducible, the identifiers of the (randomly selected) 400 podcast descriptions that have been annotated and the annotations of their contained sentences (around 2200) are made available in the project repository.

5.2 Statistics on textual content

Table 1 reports some relevant statistics about the textual content available in the Spotify Podcast dataset [5].

Podcast content is rather heterogeneous. Text distributions across different episodes are extremely variable. More specifically,

- Speech transcriptions comprise both monologues and conversations between multiple speakers.
- The speech content may cover either a small part of the episode or constitute the main substance.
- The length of both podcast descriptions and speech sentences are rather variable across different episodes.
- The length of the filtered version of the description (in terms of number of words) is approximately similar to those a single sentence.

The latter observation supports the hypothesis that the problem statement reported in Section 3 can be modelled as an extreme extractive summarization task, i.e., we set K to 1.

5.3 Statistics on audio content

Table 2 summarizes the key audio statistics. They highlight the significant variability in temporal duration across different podcasts and the wide range of loudness covered by the single podcasts. Figure 3 shows a ten-seconds excerpt of a single audio track, which reveals the presence of a considerable amount of information (b) and (c) that can be derived directly from the original waveform (a). This confirms that the audio content conveys potentially relevant and actionable knowledge.

	Episode per Show		Sentence per Episode	
	Avg #	Max #	Avg #	Max #
Train	5.77 ± 16.33	351	84.52 ± 56.97	574
Test	5.38 ± 13.71	122	78.07 ± 49.91	501
	Words per Sentence		Words per Description	
	Avg #	Max #	Avg #	Max #
Train	72.85 ± 33.34	175	61.56 ± 60.22	709
Test	75.99 ± 31.77	183	68.81 ± 54.25	461

Table 1: Statistics about the textual content extracted from the Spotify podcast dataset [5]. Minimum values are not reported here because are always equal to 1.

	Episode Duration (min)			Episode Loudness (dBFS)		
	Avg	Max	Min	Avg	Max	Min
Train	36.0 ± 23.1	304.9	0.5	-23.30 ± 4.70	-7.45	-57.63
Test	34.5 ± 20.0	89.7	1.0	-23.27 ± 5.99	-11.06	-139.00

Table 2: Statistics about the audio tracks extracted from the Spotify podcast dataset.

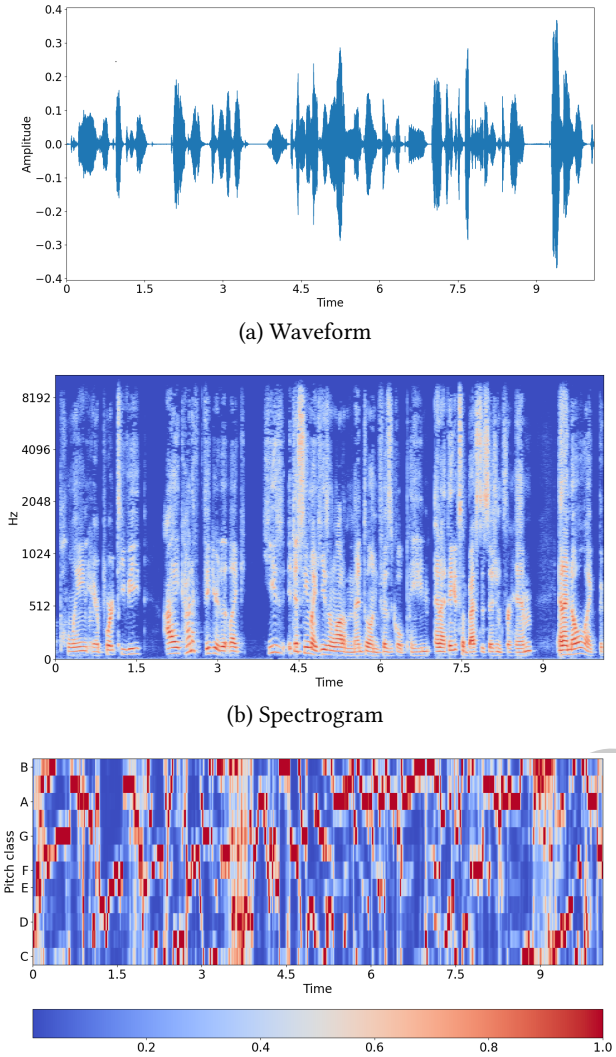
The audio extracts considered in the preliminary analyses appear to include content that is potentially actionable for text summarization.

6 EXPERIMENTAL RESULTS

We performed an extensive experimental evaluation to assess (i) the quality of the summaries generated by MATeR compared to those extracted by state-of-the-art monomodal extractive summarizers, and (ii) the impact of the two modalities considered by our approach. The following sections report the experimental design and discuss the main results, respectively.

6.1 Experimental design

We compare the automatically generated summaries with the human-generated descriptions provided by the podcast authors (hereafter also referred as *golden summaries*). The quantitative comparison



(a) Waveform

(b) Spectrogram

(c) Chroma Vector

Figure 3: Characteristics of a representative example of audio track. The waveform (a) shows the amplitude of each frame of the audio signal; the spectrogram (b) reveals the loudness of the signal at different frequencies over time; the chroma vector (c) expresses the energy of the signal for each pitch class over time.

is performed according to two complementary, standard quality indices:

- The Rouge scores [17], which measure the syntactic unit overlap between the generated and golden summaries.
- The BERT-based content similarity [24], which measures the semantic overlap between the automatically extracted and reference text snippets.

As Rouge scores we consider Rouge-1, Rouge-2, and Rouge-L, which respectively quantify the unit overlap in terms of unigrams, bigrams, and longest matching sequence [17]. To estimate the semantic similarity we compute the cosine distance between the respective Sentence-BERT encodings [24] (see Figure 2).

For the syntactical quality indices we report recall, precision, and F1-score measures achieved on the test set.

We compared MATeR with both supervised and unsupervised extractive summarization methods. As unsupervised approaches we consider CoreRank (CR) [29], TextRank (TR) [22], and TextRankBM25 (TRBM25) [2]. As supervised methods we consider HiBERT [36] and a variant of the proposed approach MATeR-textonly , where we perform ablation of the audio modality from the original architecture. Finally, we also consider the LEAD baseline method [15], which applies a naive position-based approach that simply returns the first sentence of the podcast (independently of the remaining content).

Regarding the comparison with state-of-the-art summarization methods, it is worth noticing that

- To the best of our knowledge, MATeR is the firstly proposed supervised multimodal extractive summarization approach that combines textual and acoustic features.
- The majority of the neural supervised summarization methods (e.g. [19, 38]) are unable to encode, process, and extract textual snippets longer than 512 tokens due to the inherent limitations of Transformer models. In our context, the average number of words per episode is above 5000 (e.g., in the training set, 72 words multiplied by 84 sentences) thus the number of tokens is one order of magnitude larger than the maximum threshold.
  This hinders their use in podcast summarization as the available podcast descriptions are, on average, longer (see Table 1). HiBERT [36] is, to the best of our knowledge, the most recently proposed extractive supervised summarization system that allows the processing of input sequences longer than 512 tokens.

Algorithms' configuration. We fine-tuned the pre-trained HiBERT model for 5 epochs using the parameters recommended by the respective authors. We pick the best model according to the average loss on the evaluation set.

We trained MATeR for a total of 2 epochs to minimize the MSE loss with AdamW [20] optimizer. We set the learning rate to $lr = 10^{-5}$ with a linear decay schedule.

Hardware settings. Experiments were run on a machine equipped with AMD® Ryzen 9® 3950X CPU, Nvidia® RTX 3090 GPU, 128 GB of RAM running Ubuntu 21.10.

6.2 Results

Table 3 reports the Rouge and Sentence-BERT similarity (SBERT) scores achieved by the summarization algorithms on the test set. For each considered metric the highest scores are highlighted in boldface.

The performances of the LEAD baseline are the worst. This indicates that solving the sentence ranking problem is, in general, not trivial. MATeR performs best regardless of the considered quality index. Compared to MATeR-textonly, which is the text-only ablation of MATeR, MATeR performs significantly better for all the

	Rouge-1 Precision	Rouge-1 Recall	Rouge-1 F1	Rouge-2 Precision	Rouge-2 Recall	Rouge-2 F1	Rouge-L Precision	Rouge-L Recall	Rouge-L F1	SBERT
LEAD	0.150*	0.170*	0.142*	0.014*	0.013*	0.011*	0.129*	0.147*	0.122*	0.350*
TR	0.154*	0.177*	0.147*	0.015*	0.016*	0.013*	0.133*	0.154*	0.127*	0.363*
CR	0.176*	0.179*	0.157*	0.030*	0.024*	0.023*	0.152*	0.154*	0.135*	0.418*
TRBM25	0.156*	0.203*	0.159*	0.017*	0.020*	0.015*	0.132*	0.174*	0.135*	0.414*
HiBERT	0.186*	0.219*	0.184	0.036*	0.033*	0.031*	0.162*	0.191*	0.160	0.482
MATeR-textonly	0.162*	0.168*	0.143*	0.016*	0.016*	0.013*	0.140*	0.146*	0.123*	0.348*
MATeR	**0.193**	**0.225**	**0.188**	**0.042**	**0.041**	**0.036**	**0.168**	**0.197**	**0.164**	**0.490**

Table 3: Rouge-1, Rouge-2, Rouge-L, and SBERT scores achieved by different summarization methods. Statistically significant performance improvements between MATeR and each of the other methods are starred.

considered metric. This confirms the usefulness of exploiting both input data modalities. The positive impact of the audio modality is also confirmed by the comparison against the supervised HiBERT algorithm.

Statistical significance test. We apply the paired t-test [7] at 95% significance level to compare MATeR with all the other approaches. In Table 3, the statistically significant differences between MATeR and the competitors are starred. The t-test confirms the statistical significance of the differences between MATeR and the other algorithms for the majority of the considered metrics.

Summary length. We have also analyzed the average length of the summaries to understand its impact on the quality indices achieved by the tested algorithms. The average summary length is roughly comparable for all the algorithms, e.g., 73 words for MATeR-textonly, 75 for CR, ~80 words for MATeR, HiBERT, TR, and LEAD, and 89 for TRBM25.

6.3 Summary examples

To allow a qualitative comparison between the generated summaries, Figure 4 reports the summaries generated by MATeR, HiBERT, and MATeR-textonly, respectively, from two example podcasts in which the acoustic modality plays a relevant role. We report the author's description, the textual summaries, and the chroma vectors of the selected audio summaries. It is worth noticing that the generated summaries are rather different both in terms of text and chroma vector.

The first example (see Figure 4(a)) shows that the summaries generated by MATeR and MATeR-textonly are completely different. MATeR-textonly , which tends to assign high relevance values to the initial sentences of the podcasts, selects the first sentence of the transcript. In this particular case, it seems not a good choice. Conversely, MATeR shortlists a sentence that is more informative and aligned with the description thanks to the simultaneous analysis of the acoustic features. HiBERT selects a sentence that is somehow related to the main content of the podcast as well. However, the HiBERT summary seems to be less compliant with the authors' description than those returned by MATeR.

The second summary example, reported in Figure 4(b), confirms the positive impact of the acoustic features considered by MATeR on the perceived summary quality.

7 CONCLUSIONS AND DISCUSSION

The paper explored the use of multimodal content to produce summaries of podcast episodes. An end-to-end attention-based deep learning architecture is proposed to effectively combine text and audio encodings in a multimodal fusion network. Unlike traditional text- or speech-only summarizers, MATeR leverages the multimodal information conveyed by the syncronized text-audio snippets to face the high heterogeneity of podcast contents. The summarization performance is superior to that of state-of-the-art extractive methods (e.g., [36]) on a real-world dataset.

The peculiar characteristics of the human-generated podcast descriptions has prompted the use of an extreme summarization framework, where a single sentence is representative of most of the podcast content.

As future work, we plan to extend the current transformer-based architecture, implementing a hierarchical structure to attend relevant information from multiple sentences of the same episode and/or multiple episodes of the same show. Moreover, we envisage the use of different acoustic feature extractors to leverage audio spectrogram in addition to the raw waveform.

REFERENCES

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems* 33 (2020).

[2] Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization. *arXiv preprint arXiv:1602.03606* (2016).

[3] Jingwen Bian, Yang Yang, and Tat-Seng Chua. 2013. Multimedia Summarization for Trending Topics in Microblogs. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM '13)*. Association for Computing Machinery, New York, NY, USA, 1807–1812. https://doi.org/10.1145/2505515.2505652

[4] Jingqiang Chen and Hai Zhuge. 2018. Extractive Text-Image Summarization Using Multi-Modal RNN. In *2018 14th International Conference on Semantics, Knowledge and Grids (SKG)*. 245–248. https://doi.org/10.1109/SKG.2018.00033

[5] Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 Podcasts: A Spoken English Document Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 5903–5917. https://doi.org/10.18653/v1/2020.coling-main.519

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] Thomas G. Dietterich. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* 10, 7 (1998), 1895–1923.

Author's annotated description

These horoscopes are month-ahead forecasts for each sign for Virgo Season in 2019. Virgo Season extends between August 23 - September 23. In this episode I'll take you on a tour of Virgo's zodiacal energy and explore how it manifests in the world and in each of us. Everyone has every sign in their chart, and Virgo represents amazing and important energy for each of us. This episode is a great preparation for the next 30 days of Virgo season and is also a very healing and supportive energy to check in with at any point. Listen here: Get the Virgo Season Month Ahead Extended Forecast by becoming a subscriber today!

MATeR

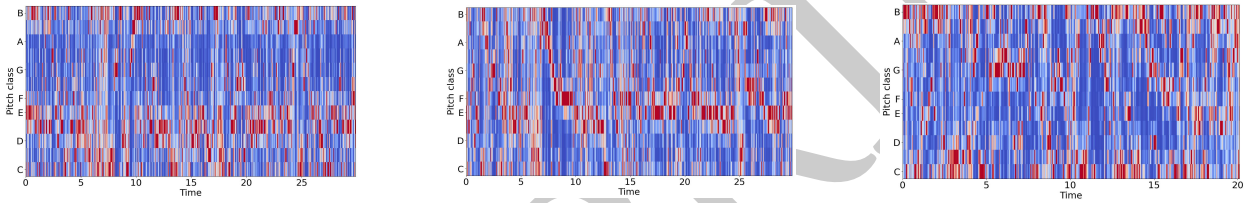
You can find out what your rising sign is by getting a free natal chart on my website embodied astrology.com in the horoscope section. If you enjoy your horoscope, please make sure to take a listen to embodied astrology for Virgo season. That's a special episode called Divinity is in the details in this episode. I take you on a tour of verbose zodiacal energy and explore how it manifests in the world and in each of us. Everyone has every sign in their chart and Virgo represents amazing and important energy for each of

HiBERT

For Peugeot season for the sign Gemini welcome Gemini to your horoscope. Please listen to the embodied astrology episode the full episode for Virgo to learn more about it. This is an important sign for you. And in that episode. I'm going to go deep into how to censor go and how to work with it somatically energetically and behaviorally and one of the reasons why I want you to listen not just because

MATeR-textonly

Apple podcasts and many other platforms the best thing or one of the best things I think about it is that you can make money from your podcasts with no minimum listener ships. It's everything you need to make a podcast in one place learn more about it and download the free anchor app or go to Anchor dot f m-- to get started.



(a) Example #1

Author's annotated description

This week I speak to Fadl Hejazi, an India based academic in Islamic sciences, currently pursuing his doctoral studies in the energy economies of the Arab world. Fadl has extensively travelled across India and he is also a regular visitor to Kashmir. In this second part podcast, more than 2-weeks since the Indian government's actions in Kashmir, I want to explore the doctrine that underpins Modi's India, known as Hindutva and its impact upon the countries Muslims. It is oft-forgotten, that after partition in 1947, Pakistan became home to just a portion of India's Muslims. Those that were left behind have been subject to state and structural disadvantage. Nevermore so than under the Hindu nationalism of Modi's tenure.

MATeR

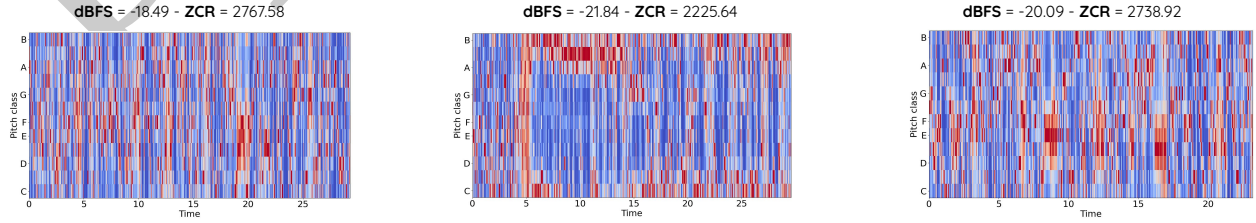
This is Mohammed Jalal and this week. I'm going to speak to further. Hey, Jersey and engine based academic in Islamic Sciences currently pursuing his doctoral studies in the energy economies of the Arab world further has extensively traveled across India, and he is also a regular visitor to Kashmir in the second part podcast more than two weeks since for Indian government's actions in Kashmir. I want to explore the doctrine that underpins mahdi's India.

HiBERT

further hejazi assalamualaikum warahmatullah and welcome to the thinking Muslim podcast walekum, Salam Allahu this much-needed discussion about the smear which today goes through a crisis a very unique crisis and unforeseen reality, which probably the ummah went through only at the time of independence of

MATeR-textonly

Says two of whom were his sons also beaten and despite one of the assailants admitting to the murder to an undercover reporter around 50 have been lynched in the last three years Alone by the so-called Cal Vigilantes and hundreds have been injured. This is the state of India home to 200 million Muslims some 14% of the population.



(b) Example #2

Figure 4: Example summaries.

https://doi.org/10.1162/089976698300017197

[8] B. Erol, D.-S. Lee, and J. Hull. 2003. Multimodal summarization of meeting recordings. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, Vol. 3. III–25. https://doi.org/10.1109/ICME.2003.1221239

[9] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal Saliency and Fusion for Movie Summarization Based on Aural, Visual, and Textual Attention. *IEEE Transactions on Multimedia* 15, 7 (2013), 1553–1568. https://doi.org/10.1109/TMM.2013.2267205

[10] Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu. 2020. Exploring wav2vec 2.0 on speaker verification and language identification. *CoRR* abs/2012.06185 (2020). arXiv:2012.06185 https://arxiv.org/abs/2012.06185

[11] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori. 2004. Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing* 12, 4 (2004), 401–408. https://doi.org/10.1109/TSA.2004.828699

[12] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

[13] Rosie Jones. 2020. The New TREC Track on Podcast Search and Summarization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 7. https://doi.org/10.1145/3397271.3402431

[14] Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth JF Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. 2021. Trec 2020 podcasts track overview. *arXiv preprint arXiv:2103.15953* (2021).

[15] Rosie Jones, Ben Carterette, Ann Clifton, Jussi Karlgren, Aasish Pappu, Sravana Reddy, Yongze Yu, Maria Eskevich, and Gareth J. F. Jones. 2020. TREC 2020 Podcasts Track Overview. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020 (NIST Special Publication)*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. 1266. National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.P.pdf

[16] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal Summarization for Asynchronous Collection of Text, Image, Audio and Video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 1092–1102. https://doi.org/10.18653/v1/D17-1114

[17] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[18] Tzu-En Liu, Shih-Hung Liu, and Berlin Chen. 2019. A Hierarchical Neural Summarization Framework for Spoken Documents. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7185–7189. https://doi.org/10.1109/ICASSP.2019.8683758

[19] Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. 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)*. 3730–3740.

[20] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=Bkg6RiCqY7

[21] Potsawee Manakul and Mark Gales. 2021. Long-Span Summarization via Local Attention and Content Selection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 6026–6041. https://doi.org/10.18653/v1/2021.acl-long.470

[22] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain, 404–411. https://aclanthology.org/W04-3252

[23] Sravana Reddy, Yongze Yu, Aasish Pappu, Aswin Sivaraman, Rezvaneh Rezapour, and Rosie Jones. 2021. Detecting Extraneous Content in Podcasts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 1166–1173. https://doi.org/10.18653/v1/2021.eacl-main.99

[24] Nils Reimers, Iryna Gurevych, Nils Reimers, Iryna Gurevych, Nandan Thakur, Nils Reimers, Johannes Daxenberger, Iryna Gurevych, Nils Reimers, Iryna Gurevych, et al. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[25] Rezvaneh Rezapour, Sravana Reddy, Ann Clifton, and Rosie Jones. [n. d.]. Spotify at TREC 2020: Genre-Aware Abstractive Podcast Summarization. ([n. d.]).

[26] Rezvaneh Rezapour, Sravana Reddy, Ann Clifton, and Rosie Jones. 2021. Spotify at TREC 2020: Genre-Aware Abstractive Podcast Summarization. *CoRR* abs/2104.03343 (2021). arXiv:2104.03343 https://arxiv.org/abs/2104.03343

[27] Manos Schinas, Symeon Papadopoulos, Georgios Petkos, Yiannis Kompatsiaris, and Pericles A. Mitkas. 2015. Multimodal Graph-Based Event Detection and Summarization in Social Media Streams. In *Proceedings of the 23rd ACM International Conference on Multimedia (MM '15)*. Association for Computing Machinery, New York, NY, USA, 189–192. https://doi.org/10.1145/2733373.2809933

[28] Kaiqiang Song, Chen Li, Xiaoyang Wang, Dong Yu, and Fei Liu. 2020. Automatic summarization of open-domain podcast episodes. *arXiv preprint arXiv:2011.04132* (2020).

[29] Antoine Tixier, Polykarpos Meladianos, and Michalis Vazirgiannis. 2017. Combining Graph Degeneracy and Submodularity for Unsupervised Extractive Summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*. Association for Computational Linguistics, Copenhagen, Denmark, 48–58. https://doi.org/10.18653/v1/W17-4507

[30] Nik Vaessen and David A van Leeuwen. 2021. Fine-tuning wav2vec2 for speaker recognition. *arXiv preprint arXiv:2109.15053* (2021).

[31] Aneesh Vartakavi and Amanmeet Garg. 2020. PodSumm–Podcast Audio Summarization. *arXiv preprint arXiv:2009.10315* (2020).

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[33] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5005–5013. https://doi.org/10.1109/CVPR.2016.541

[34] Shasha Xie, Hui Lin, and Yang Liu. 2010. Semi-supervised extractive speech summarization via co-training algorithm. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura (Eds.). ISCA, 2522–2525. http://www.isca-speech.org/archive/interspeech_2010/i10_2522.html

[35] Chenxi Zhang, Zijian Zhang, Jiangfeng Li, Qin Liu, and Hongming Zhu. 2021. CtnR: Compress-then-Reconstruct Approach for Multimodal Abstractive Summarization. In *2021 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN52387.2021.9534082

[36] Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5059–5069. https://doi.org/10.18653/v1/P19-1499

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

[38] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2020. Extractive Summarization as Text Matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6197–6208.

[39] Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A Hierarchical Network for Abstractive Meeting Summarization with Cross-Domain Pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.), Vol. EMNLP 2020. Association for Computational Linguistics, 194–203. https://doi.org/10.18653/v1/2020.findings-emnlp.19