gracenote
A NIELSEN COMPANY

Audio Summarization for Podcasts

Aneesh Vartakavi, Amanmeet Garg, and Zafar Rafii

Gracenote, Emeryville, CA, USA firstname.lastname@nielsen.com

29th EUSIPCO
European Signal Processing Conference
DUBLIN // IRELAND
23–27 AUGUST 2021

Summary

- *PodSumm*: first system to automatically generate extractive audio summaries for podcasts, for listeners to preview episodes [1].
- Uses ASR on the audio and extractive text summarization on the transcript.
- Created internal dataset of summaries from podcasts and used it to fine-tune a Transformer-based summarization model [2].
- Good performance for podcast summarization with ROUGE-(1/2/L) F-scores of 0.63/0.53/0.63 [3].
- Examples: `https://github.com/aneeshvartakavi/podsumm`

PodSumm Architecture

❶**Automatic speech recognition (ASR)**: Generate a transcript of the podcast episode using AWS Transcribe (`https://aws.amazon.com/transcribe/`).

❷**Text processing**: Parse the transcript into individual sentences, with their time offsets in the audio, using spaCy (`https://spacy.io/usage/linguistic-features#sbd`).

❸**Text summarization**: Generate a text summary by selecting the relevant sentences, using a recent BERT-based summarization model (PreSumm) [2] fine-tuned on our own dataset of podcast summaries.

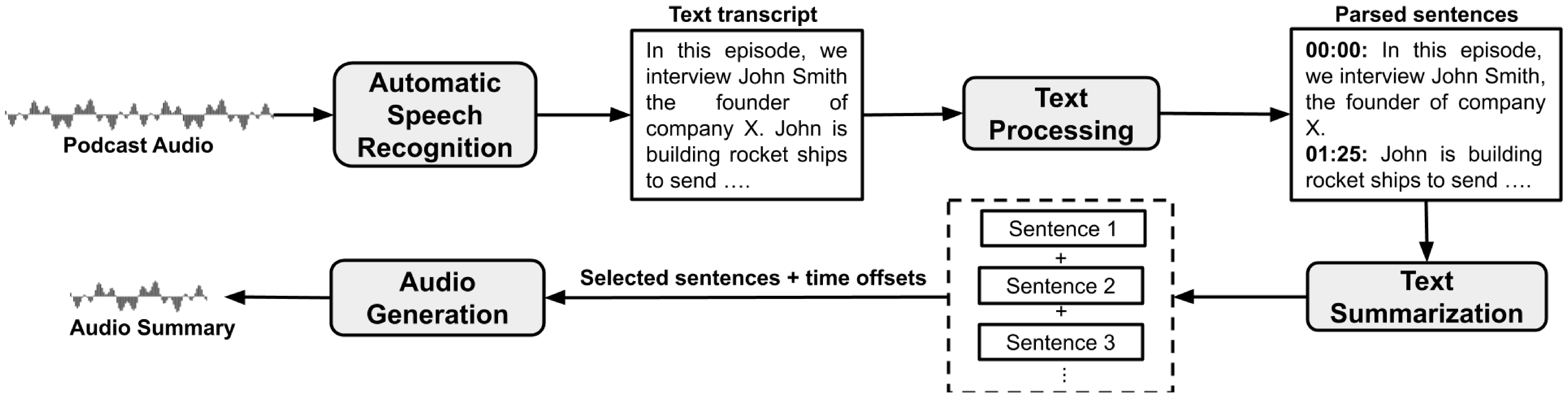
❹**Audio generation**: Derive the audio summary by using the time offsets of the selected sentences in the podcast and stitch them together.

Dataset Creation

- Selected 19 podcast series, 309 podcast episodes, 188 hours of audio.
- Built a tool to present annotators with the transcript of an episode and let them select the sentences that best represent a summary.
- Had 17 annotators who selected about 15 sentences per summary.

Model Training

- Fine-tuned PreSumm model (trained on the CNN/DailyMail dataset) on our podcast dataset for 10,000 steps as in [2].
- Selected the top-k sentences to create the final summary.
- Trained/tested model on 80%/20% split, with 5-fold cross-validation.
- Shuffled repetitive segments between podcasts as data augmentation.



Podcast Audio

Automatic Speech Recognition

Text transcript

In this episode, we interview John Smith the founder of company X. John is building rocket ships to send ....

Text Processing

Parsed sentences

00:00: In this episode, we interview John Smith, the founder of company X.
01:25: John is building rocket ships to send ....

Text Summarization

Sentence 1
+
Sentence 2
+
Sentence 3
⋮

Selected sentences + time offsets

Audio Generation

Audio Summary

Figure 1:Overview of the PodSumm system.

Results

- Reported the F-scores for the ROUGE-(1/2/L) metric [3], which measures the overlap of n-grams between predicted and reference summaries (n = 1, 2, Longest common subsequence).
- Compared to baseline LEAD-k which selects the first k sentences as the summary, and PreSumm (k=12) not fine-tuned on our dataset.
- Obtained the best results with PreSumm fine-tuned (FT) and with data augmentation (Aug) for k = 12 sentences.
- Showed that podcast summarization can be done in the text domain.

Metric	R-1 F1	R-2 F1	R-L F1
LEAD-k (baseline)			
$k = 5$	0.28 (0.02)	0.17 (0.03)	0.27 (0.02)
$k = 9$	0.40 (0.03)	0.26 (0.04)	0.39 (0.03)
$k = 12$	0.47 (0.03)	0.32 (0.03)	0.46 (0.02)
$k = 15$	**0.52 (0.03)**	**0.39 (0.04)**	**0.51(0.03)**
PreSumm ($k = 12$)			
No FT	0.53 (0.02)	0.38 (0.02)	0.52 (0.02)
FT	0.63 (0.03)	0.51 (0.03)	0.62 (0.03)
FT + Aug	**0.64 (0.02)**	**0.53 (0.03)**	**0.63 (0.02)**
PreSumm (FT + Aug)			
$k = 5$	0.56 (0.03)	0.46 (0.04)	0.55 (0.03)
$k = 9$	0.63 (0.02)	0.52 (0.03)	0.62 (0.02)
$k = 12$	**0.64 (0.02)**	**0.53 (0.03)**	**0.63 (0.02)**
$k = 15$	0.63 (0.02)	0.53 (0.03)	0.62 (0.02)

Table 1:Mean (and standard deviation) of the ROUGE-(1/2/L) F-scores, for the baseline, PreSumm with $k = 12$, and PreSumm (FT + Aug) with $k \in (5, 9, 12, 15)$, for a 5-fold cross validation. Higher scores are better. Bold values are the highest.

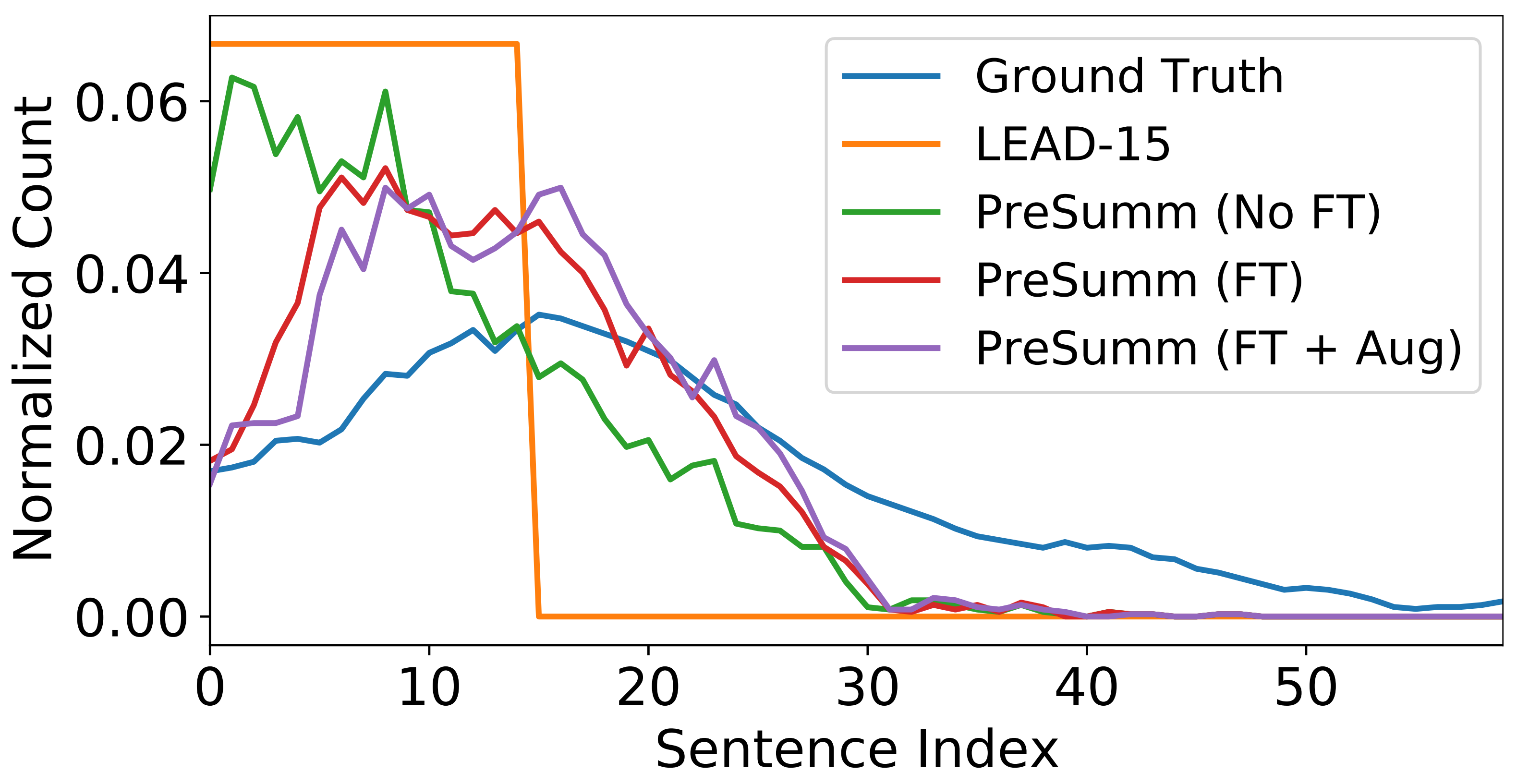


Figure 2:Distribution of the sentence indices corresponding to the best summary candidates, over all the podcast episodes in our dataset, for the different methods. As the proposed model improves, the ROUGE F-score increases, and the distribution shifts closer towards the ground truth.

References

[1] Aneesh Vartakavi and Amanmeet Garg, "PodSumm: Podcast audio summarization," PodRecs: The Workshop on Podcast Recommendations, September 25 2020.

[2] Yang Liu and Mirella Lapata, "Text summarization with pretrained encoders," in *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, November 3–7 2019.

[3] Chin-Yew Lin, "ROUGE: A package for automatic evaluation of summaries," in *Workshop on Text Summarization Branches Out*, Barcelona, Spain, July 25-26 2004.