Voice-Driven Content Summarization

Arnav Sharma[a], Himanshu[a]

[a]*Department of Computer Science and Engineering, Indian Institute of Technology ,Jodhpur, Karwar, 342030, Rajasthan, India*

Abstract

Voice-driven content summarization is an emerging task within speech processing that focuses on converting long, spoken audio recordings into concise, coherent text summaries. This process holds immense value across a wide range of real-world applications, including enhancing accessibility, improving information retrieval systems, and reducing the time and effort involved in processing large amounts of spoken data. For instance, in sectors like healthcare, voice summarization can alleviate the burden of clinical documentation, while in media and business environments, it can significantly streamline the management of meetings, interviews, and lectures.

The task of speech summarization can be approached through various methodologies, which typically fall under one of four key architectures: sentence extraction and compaction, feature extraction with classification or ranking, sentence compression, and language modeling. These methods aim to create summaries that are not only accurate but also contextually meaningful, ensuring that the essence of the original speech is preserved. However, challenges remain in terms of balancing summarization quality and efficiency, particularly in domains where manual annotation is expensive or impractical.

This report delves into the strengths and limitations of state-of-the-art models for voice-driven content summarization, drawing insights from the recent literature, including a scoping review of 110 studies. While supervised approaches like Hidden Markov Support Vector Machines (SVMs) and Conditional Random Fields (CRFs) have demonstrated superior performance, they are hindered by the need for large annotated datasets. Consequently, recent trends in unsupervised techniques, including deep learning models combined with unigram language modeling, are gaining momentum due to their ability to perform without extensive labeled data.

Furthermore, this report evaluates the metrics used to assess summarization quality, highlighting their respective advantages and drawbacks. Finally, it identifies open problems and opportunities in the field, such as the need for more robust unsupervised learning methods, the integration of multimodal data, and the enhancement of summarization algorithms for more diverse and dynamic speech content. Through this comprehensive analysis, the report aims to shed light on the current landscape of voice-driven content summarization and its future directions.

1. Introduction

Voice-driven content summarization is a task within the field of speech processing that involves transforming lengthy audio recordings into concise, easily understandable text summaries. The goal is not just to provide a verbatim transcript, but to identify and extract the most important information, distilling it into a more digestible form that highlights key points while removing extraneous details such as pauses, repetitions, and informal speech patterns. This type of summarization is crucial in environments where processing large volumes of spoken data is necessary,

but time and resources are limited. By summarizing spoken content, it becomes easier to access relevant information quickly, improving both efficiency and productivity.

The task of speech summarization has become more significant in recent years due to several factors. Advances in Automatic Speech Recognition (ASR) technology, improvements in audio capture quality, and the increasing use of natural language processing (NLP) models have contributed to the growing interest in this area. As a result, speech summarization has found a wide range of applications in various sectors. In media, it is used to summarize broadcast news or interviews. In corporate environments, it is applied to summarize meetings and discussions, allowing teams to quickly review key decisions and action points. In the healthcare industry, voice-driven summarization has the potential to reduce the administrative burden on clinicians by automatically generating clinical records from doctor-patient conversations.

Speech summarization systems generally operate through two main approaches: extractive and abstractive. Extractive summarization focuses on identifying the most important segments from the audio and directly concatenating them to form a coherent summary. Abstractive summarization, on the other hand, generates new sentences that convey the essence of the original content, rephrasing and condensing information where necessary. Abstractive summarization is more challenging, as it requires the system to understand the underlying meaning of the speech and generate new text, which demands advanced capabilities in natural language generation.

Current research has evolved significantly from earlier methods, particularly with the advent of deep learning techniques. While previous literature primarily emphasized traditional methods such as sentence extraction and word-based sentence compaction, recent advancements leverage neural networks and other machine learning models to improve summarization accuracy and fluency. However, while the results have been promising, challenges remain, including the need for large, annotated training datasets and the complexity of creating accurate summaries from diverse, real-world speech content.

In the context of voice-driven content summarization, the system's ability to process long, continuous speech and generate concise, coherent summaries is of critical importance. This task has a broad range of practical applications, including improving efficiency in customer service, facilitating learning through online education, and easing the documentation processes in healthcare. Despite the progress made in summarization techniques, there are still open challenges, particularly in handling spontaneous, unstructured speech and maintaining the balance between accuracy and brevity in the generated summaries.

This report aims to provide a detailed analysis of the state-of-the-art models for voice-driven content summarization, exploring both the extractive and abstractive approaches. We will examine the strengths and limitations of these models, the metrics used to evaluate their performance, and the opportunities and challenges in advancing this field further.

2. Speech Recognition Approaches [1]

Speech recognition, also known as Automatic Speech Recognition (ASR), is the process of converting spoken language into text. Over the years, various approaches have been developed to improve accuracy and adaptability in different environments. The three primary approaches to ASR are the Acoustic-Phonetic Approach, the Pattern Recognition Approach, and the Artificial Intelligence (AI) Approach. Figure 2 provides an overview of these methods and their respective subcategories.

2.1. Acoustic-Phonetic Approach

The acoustic-phonetic method [2] is based on the assumption that spoken language consists of distinct phonetic units characterized by a set of acoustic features. However, these features vary due to speaker differences and background noise, making the task more complex. The approach involves the following steps:

- Speech Spectrum Analysis – Extracting acoustic features from speech signals
- Feature Extraction – Identifying relevant phonetic components.
- Segmentation – Dividing speech into phonetic units.
- Labeling – Assigning phonetic labels to segments.

Although this approach provides a theoretical basis for ASR, it has not been widely adopted due to the complexity of mapping phonetic units to speech variations.

2.2. Pattern Recognition Approach

Pattern recognition [3]is the most widely used approach in ASR. It involves extracting patterns from speech signals and classifying them into known linguistic categories. This approach consists of the following methodologies:

1. **Template-Based Approach** : This method [4] relies on a database of reference speech patterns. An unknown speech input is compared against stored templates, and the best-matching pattern is selected. While effective, this approach struggles with speaker variability and large vocabulary applications.
2. **Stochastic Approach**: Stochastic modeling leverages probabilistic frameworks to handle speech variations. The Hidden Markov Model (HMM) is the most commonly used stochastic model in ASR. HMMs effectively model temporal dependencies in speech signals, making them suitable for continuous speech recognition tasks.
3. **Dynamic Time Warping (DTW)**:DTW [5]is used to measure the similarity between two sequences of different lengths. This technique allows flexible alignment of speech patterns, making it useful for handling variations in speech speed and pronunciation.
4. **Vector Quantization (VQ)**: VQ creates a codebook of speech feature vectors for speaker-specific recognition. While computationally efficient, VQ lacks temporal information, limiting its use in continuous speech recognition.

2.3. Artificial Intelligence (AI) Approach

The AI-based approach [5] integrates principles from both acoustic-phonetic and pattern recognition methodologies. It leverages machine learning techniques to improve speech recognition accuracy and generalization.

1. **Knowledge-Based Approach** : This method incorporates linguistic, phonetic, and spectrogram-based knowledge to analyze speech. While theoretically promising, the integration of expert knowledge remains a challenge.
2. **Connectionist Approach (Artificial Neural Networks - ANN)**:Neural networks, particularly deep learning architectures, have significantly advanced ASR performance. Multilayer neural networks learn complex speech patterns and classify phonetic units. This approach is widely used in modern ASR systems due to its robustness.

3. **Support Vector Machine (SVM)**:SVM is a powerful classification algorithm that separates speech data using hyperplanes. While effective for fixed-length data, variable-length speech inputs require transformation before classification.

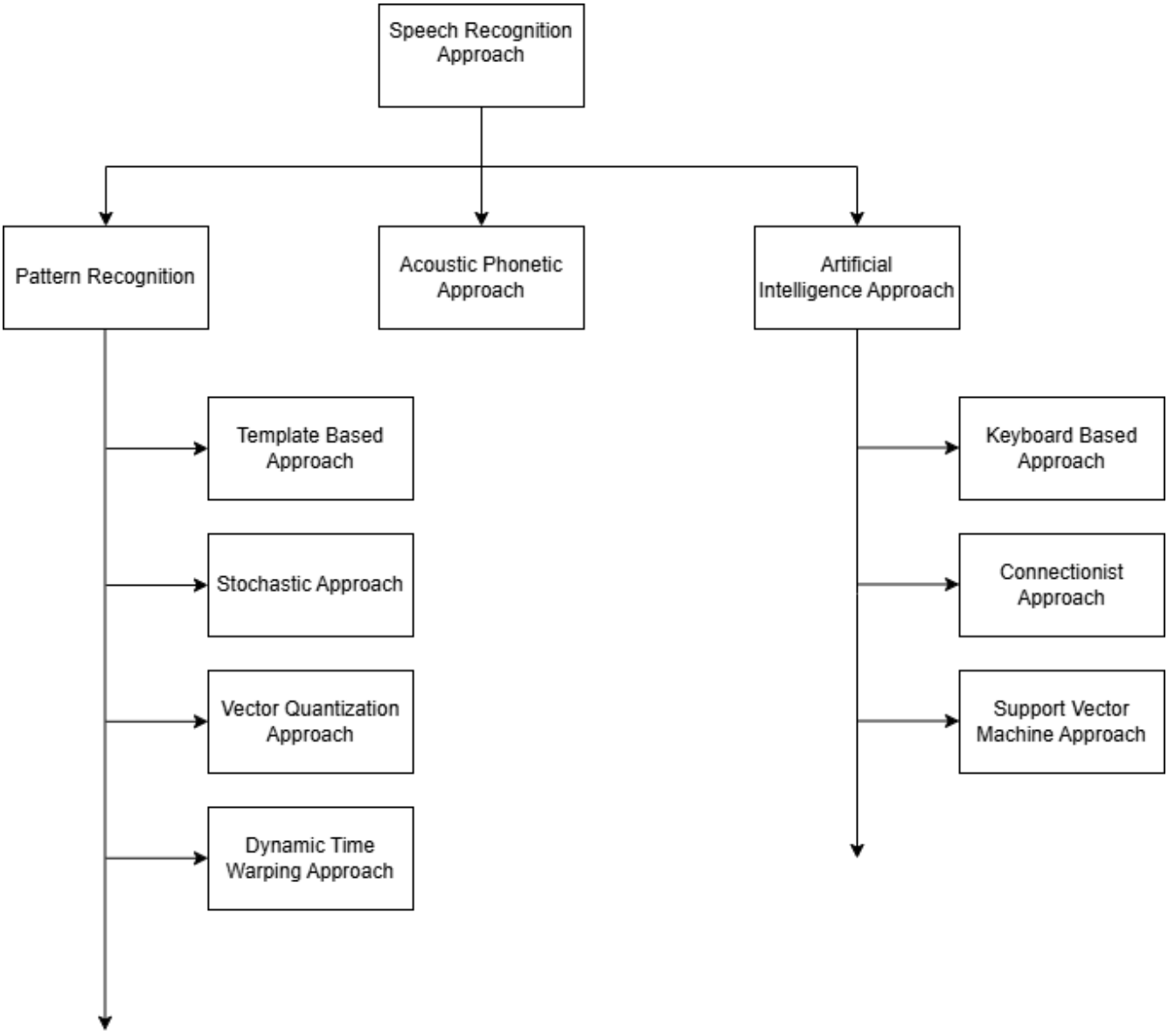

Figure 1: Speech Recognition Approaches

3. Evaluation Metrics

Summarization quality is highly subjective—what one user finds informative or coherent may differ from another's perception. Early researchers often relied on human evaluations, such as Likert-scale ratings or qualitative comparisons, because it was difficult to capture the nuances of summary quality using a single, well-accepted automatic metric. However, as the field evolved, several quantitative metrics were introduced and have since been used in the literature.

3.1. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE [6] is a set of metrics commonly used to evaluate automatic summarization and machine translation. It measures the overlap between the generated summary and reference summaries by comparing n-grams, word sequences, and word pairs. The primary variants include: ROUGE-1, Measures the overlap of unigrams (single words) between the system and reference summaries. ROUGE-2, Focuses on bigram (two-word sequence) overlap. ROUGE-L, Considers the longest common subsequence, capturing sentence-level structure similarity. These metrics primarily emphasize recall, assessing how much of the reference summary's content is captured by the generated summary.

$$\text{ROUGE-N} = \frac{\sum_{\text{gram}_n \in \{\text{Reference Summaries}\}} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{\text{gram}_n \in \{\text{Reference Summaries}\}} \text{Count}(\text{gram}_n)}$$

3.2. METEOR (Metric for Evaluation of Translation with Explicit ORdering)

METEOR addresses some limitations of precision-focused metrics by incorporating recall and considering linguistic features. It aligns words in the candidate and reference texts using exact matches, stemming (matching words with the same root), and synonyms. The final score is a harmonic mean of precision and recall, with recall often weighted higher, and includes a penalty for fragmented matches to account for fluency.

$$\text{METEOR} = F_{\text{mean}} \times (1 - \text{Penalty})$$

$$F_{\text{mean}} = \frac{10 \times P \times R}{9 \times P + R}$$

$$\text{Penalty} = 0.5 \times \left(\frac{c}{m}\right)^3$$

3.3. BERT Score

BERTScore is a more recent metric that leverages contextual embeddings from the BERT language model to evaluate text generation. Instead of relying solely on surface-level token matches, it computes the cosine similarity between embeddings of the generated and reference texts, capturing deeper semantic similarities. This approach allows BERTScore to account for paraphrasing and variations in wording that still convey the same meaning.

3.4. Precision

The proportion of relevant instances among the retrieved instances. It reflects how much of the generated summary is relevant.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

3.5. Recall

The proportion of relevant instances that have been retrieved over the total amount of relevant instances. It indicates how much of the reference summary's content is captured.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

3.6. F1-Score

The harmonic mean of precision and recall, providing a single metric that balances both concerns.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. Literature Review

4.1. Automatic Speech Summarization Using Dynamic Programming and Dependency Structures

The [7] proposed Automatic Speech Summarization (ASS) system transcribes speech using LVCSR and applies Dynamic Programming (DP) to extract words that maximize a summarization score, incorporating word significance (I), confidence (C), linguistic correctness (L), and word concatenation (Tr) for coherence. Evaluated on Japanese broadcast news, it achieved 70% compression for single utterances and 30% for multi-utterances, using word networks for accuracy assessment. The model effectively handles recognition errors and ensures fluency, but is language-dependent, computationally expensive, and lacks neural network integration. Unlike modern transformer-based SOTA models (e.g., BERT, GPT), it is not end-to-end or multilingual, though its dependency structures could enhance deep learning approaches.

4.2. Speech-to-Text and Speech-to-Speech Summarization of Spontaneous Speech

This paper [8] introduces a two-stage method for summarizing spontaneous speech, focusing on both speech-to-text and speech-to-speech summarization. For speech-to-text, the approach first extracts important sentences based on linguistic likelihood, word significance, and confidence scores, then further compacts them using dynamic programming for coherence. Speech-to-speech summarization explores extracting and concatenating sentence units, word units, and between-filler units, ensuring natural transitions. Evaluated on the Corpus of Spontaneous Japanese (CSJ), the method outperforms random selection, particularly at a 50% summarization ratio. While sentence units yield the most natural speech summaries, word-based summarization suffers from unnatural concatenation. Strengths include adaptability to spontaneous speech and a novel evaluation framework, though language dependency and limited dataset size remain constraints. Future work aims at broader datasets and improved unit extraction techniques.

4.3. Hybrid Speech-to-Text and Text Summarization Approach

This paper [9] presents a hybrid approach integrating speech recognition with text summarization to enhance documentation efficiency in applications such as lecture notes generation and document processing. Speech is transcribed using Google Speech API, and a custom punctuation insertion algorithm improves text segmentation for better summarization. The summarization model leverages sentence tokenization and word frequency analysis with a custom ranking method, outperforming Gensim in efficiency and coherence. Experimental results show improved speech recognition speed and structured summaries due to pre-processing. While effective, the approach relies on external APIs and may need optimization for diverse speech variations.

4.4. PodSumm: Automated Podcast Summarization

Podcasts [10] pose challenges for content discovery, as creator-provided descriptions often lack key subjective details like narration style and production quality. PodSumm addresses this by generating audio summaries through a two-step process: AWS Transcribe for speech-to-text conversion and a fine-tuned PreSumm (BERT-based) model for extractive summarization. A custom

dataset of 188 hours from 19 podcast series was created, with human-annotated key sentences for training. Fine-tuning and data augmentation improved summarization quality, achieving strong ROUGE scores (0.63/0.53/0.63). PodSumm effectively previews podcast content, enabling efficient discovery and paving the way for further research in audio summarization.

4.5. Adaptive Beam Search for On-Device Abstractive Summarization

Adaptive Beam Search (ABS) [11] prioritizes source-specific keywords dynamically, improving summary relevance for SMS, voice messages, and documents. Knowledge Distillation (KD) reduces model size by 30.9% while maintaining performance, enabling faster inference and lower memory usage. A Bigram Language Model (LM) enhances grammatical accuracy by prioritizing meaningful word pairs. A novel scoring strategy selects the best summary hypothesis, balancing keyword frequency and fluency. This lightweight, efficient model is optimized for on-device deployment, eliminating reliance on cloud processing.

4.6. Voice-Based Summary Generation using LSTM

Utilizes an LSTM-based [12] sequence-to-sequence model for summarizing long voice recordings. Voice-to-text conversion is performed using Python's speech_recognition library, with audio split into smaller chunks for better transcription accuracy. The model is trained on a news summary dataset, employing encoder-decoder LSTM architecture where the input text is processed sequentially, and summaries are generated word-by-word. Special tokens ¡start¿ and ¡end¿ help structure summaries. The approach effectively handles long recordings, making it suitable for meeting summaries, interviews, and large conversations.

4.7. Towards End-to-End Speech-to-Text Summarization

This work explores [13] both cascade and end-to-end (E2E) approaches for speech-to-text (S2T) abstractive summarization of broadcast news, where the cascade model transcribes speech using ASR before summarization, while the E2E model directly processes speech features via a self-supervised pre-trained model and transfer learning from a text-to-text (T2T) summarizer. While both models outperform extractive baselines, the E2E model lags behind the cascade system due to limited training data, particularly affecting the cross-modal adapter's performance. However, the study narrows the performance gap and highlights the potential of direct speech-to-summary generation, emphasizing the need for richer training data to enhance E2E summarization.

4.8. An End-to-End Speech Summarization Using Large Language Model

This study [14] presents an end-to-end speech summarization model that integrates a large language model (LLM) with a Q-Former module to directly generate text summaries from speech features, surpassing traditional cascaded models and performing competitively on the How2 dataset. Using a multi-stage training approach, the model employs ASR and text summarization (TSum) as auxiliary tasks, leveraging curriculum learning to align feature spaces and enhance cross-modal mapping. Results show strong performance across ROUGE, METEOR, and BERTScore metrics, with the model effectively mitigating ASR error propagation and bridging the speech-text modality gap, demonstrating the potential of LLMs for direct speech summarization.

5. Research Gap

Despite significant advancements in speech-to-text abstractive summarization, several research gaps remain unaddressed, limiting the effectiveness, generalizability, and real-world deployment of current models.

5.1. Error Propagation in Cascade Systems

Cascade models rely on automatic speech recognition (ASR) before applying text summarization, making them susceptible to ASR errors. These errors degrade summarization quality, particularly in noisy environments or for low-resource languages. While end-to-end (E2E) models attempt to mitigate this issue, they still struggle with performance gaps compared to cascaded approaches. More robust techniques are needed to minimize ASR-induced degradation in the summarization process.

5.2. Limited Training Data for End-to-End Models

E2E summarization models require large-scale speech-text datasets, yet most existing corpora (e.g., MLSUM, How2) are optimized for text-based summarization rather than direct speech-to-summary tasks. The lack of high-quality paired speech-summary datasets, especially for multilingual or domain-specific applications, restricts the model's ability to generalize across various contexts. Data augmentation, semi-supervised learning, and synthetic data generation remain underexplored solutions.

5.3. Cross-Modal Feature Alignment Challenges

Aligning speech representations with text-based summarization remains a significant challenge. Most E2E models rely on self-supervised speech encoders, but there is still a gap in fully capturing non-verbal cues (e.g., speaker intent, prosody, and emphasis) that could enhance summary coherence. The integration of advanced multi-modal learning techniques, such as vision-language pretraining strategies, could improve speech-to-summary mappings.

5.4. Domain Adaptability and Generalization

Current models often struggle to generalize across diverse domains, such as legal, medical, and financial speech summarization. Most SOTA models are trained on news datasets, limiting their effectiveness in scenarios requiring domain-specific vocabulary and structure. Research into domain-adaptive summarization techniques, few-shot learning, and retrieval-augmented summarization could enhance applicability.

5.5. Computational Constraints for On-Device Deployment

While some models propose lightweight adaptations, high-performing speech summarization models remain computationally expensive, making real-time, on-device summarization challenging. Efficient knowledge distillation, quantization, and edge-computing-friendly architectures need further exploration to balance performance with computational efficiency.

5.6. Evaluation Metrics for Speech Summarization

Most studies evaluate summarization using text-based metrics like ROUGE, METEOR, and BERTScore, which may not fully capture the quality of summaries generated directly from speech. Metrics that consider speech-specific aspects, such as semantic preservation, prosodic relevance, and coherence across modalities, are still underdeveloped. Research into human-centric and speech-aware evaluation frameworks is necessary.

References

[1] T. Kumar, M. Mahrishi, G. Meena, A comprehensive review of recent automatic speech summarization and keyword identification techniques, Artificial Intelligence in Industrial Applications: Approaches to Solve the Intrinsic Industrial Optimization Problems (2022) 111–126.

[2] S. J. Arora, R. P. Singh, Automatic speech recognition: a review, International Journal of Computer Applications 60 (9) (2012).

[3] B.-H. Juang, W. Chou, C.-H. Lee, Statistical and discriminative methods for speech recognition, in: Automatic Speech and Speaker Recognition: Advanced Topics, Springer, 1996, pp. 109–132.

[4] M. Sharma, K. K. Sarma, Soft-computational techniques and spectro-temporal features for telephonic speech recognition: an overview and review of current state of the art, Artificial Intelligence: Concepts, Methodologies, Tools, and Applications (2017) 1591–1619.

[5] M. Anusuya, S. K. Katti, Speech recognition by machine, a review, arXiv preprint arXiv:1001.2267 (2010).

[6] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81.
URL https://aclanthology.org/W04-1013/

[7] C. Hori, S. Furui, Advances in automatic speech summarization, RDM 80 (2001) 100.

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

[9] A. Vinnarasu, D. V. Jose, Speech to text conversion and summarization for effective understanding and documentation, International Journal of Electrical and Computer Engineering (IJECE) 9 (5) (2019) 3642–3648.

[10] A. Vartakavi, A. Garg, Podsumm–podcast audio summarization, arXiv preprint arXiv:2009.10315 (2020).

[11] B. Harichandana, S. Kumar, Adaptive beam search to enhance on-device abstractive summarization, in: 2021 IEEE 18th India Council International Conference (INDICON), IEEE, 2021, pp. 1–6.

[12] R. Sharma, N. Varshney, M. Sharma, R. Paliwal, Voice based summary generation using lstm, Int J Res Appl Sci Eng Technol (IJRASET) 10 (06) (2022) 46–51.

[13] R. Monteiro, D. Pernes, Towards end-to-end speech-to-text summarization, in: K. Ekštein, F. Pártl, M. Konopík (Eds.), Text, Speech, and Dialogue, Springer Nature Switzerland, Cham, 2023, pp. 304–316.

[14] H. Shang, Z. Li, J. Guo, S. Li, Z. Rao, Y. Luo, D. Wei, H. Yang, An end-to-end speech summarization using large language model, arXiv preprint arXiv:2407.02005 (2024).