International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 05 | May – 2026 DOI:10.55041/ISJEM07344
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Meri Awaaz Hi Pehchan Hai: A Survey on Multilingual Podcast Processing System

Dr. Vijayalaxmi Mekali, Arpita Rathod, Bhagyashree, Gagana Poojari, Gayana V
Department of Computer Science and Engineering
K. S. Institute of Technology, Bengaluru – 560109, India vijayalaxmimekali@ksit.edu.in

***Abstract*—**Podcast consumption has increased rapidly in re-cent years, making podcasts an important source of education, entertainment, and information sharing. In multilingual countries like India, audiences prefer podcast content in regional languages such as Kannada, Hindi, and Telugu. Processing and summariz-ing multilingual podcasts remains challenging because of regional accents, code-switching, dialectal variations, and long audio content. The survey reviews multilingual podcast processing and summarization techniques, including Automatic Speech Recog-nition (ASR), Natural Language Processing (NLP), extractive and abstractive summarization, machine translation, and Text-to-Speech systems. The study also examines recent advancements in transformer-based models for low-resource Indian languages and discusses datasets, evaluation metrics, research challenges, and future improvements in multilingual podcast summarization.

Index Terms—Keywords: Podcast Summarization, Multilin-gual Automatic Speech Recognition (ASR),Natural Language Processing (NLP), Indian Languages, Transformer Models.

I. INTRODUCTION

The popularity of podcasts has increased significantly in recent years due to the rapid growth of digital media plat-forms and audio streaming services. Podcasts are now widely used for education, news broadcasting, storytelling, technical discussions, business communication, and entertainment purposes. In India, the expansion of affordable internet services and smartphone accessibility has encouraged a large number of users to consume podcast content in regional languages such as Kannada, Hindi, Telugu, Tamil, and Marathi [1, 2].

Regional language podcasts provide users with easier access to information in their native language, making digital content more inclusive and accessible. However, podcast episodes are generally lengthy and require continuous listening, which makes information extraction time-consuming for listeners. Because of this, podcast summarization has become an important research area in Natural Language Processing (NLP) and Artificial Intelligence (AI) [1].

Most existing podcast summarization systems are designed mainly for English-language datasets and often show limited performance when applied to multilingual Indian content. Indian podcasts commonly include code-mixed speech where speakers switch between English and regional languages within the same conversation. In addition, differences in dialects, pronunciation patterns, speaking styles, and background noise make multilingual podcast processing more challenging [2, 3].

To address these challenges, advanced technologies such as Automatic Speech Recognition (ASR), transformer-based NLP models, machine translation systems, and Text-to-Speech (TTS) frameworks are increasingly being adopted. Models including Whisper, BERT, BART, mBART, and mT5 have demonstrated effective performance in multilingual speech transcription and text summarization tasks [3, 4, 5]. These models support the conversion of spoken podcast audio into textual transcripts and help generate concise summaries while preserving the important contextual meaning of the original content.

Research related to multilingual podcast summarization mainly focuses on improving transcription accuracy, handling code-mixed speech, supporting low-resource Indian languages, and generating context-aware summaries. The availability of multilingual transformer models and Indian language NLP frameworks has further improved the feasibility of developing efficient regional language podcast summarization systems [6]. This survey paper examines recent developments in multilingual podcast processing systems for Indian regional languages including English, Kannada, Hindi, and Telugu. The study reviews major summarization techniques, ASR technologies, transformer-based models, datasets, and existing research challenges associated with multilingual podcast summarization [1, 2].

II. RELATED WORKS

A. Podcast Summarization for Regional Languages

A survey conducted by Lokhande et al. examined podcast summarization systems designed for regional language processing [1]. The study discussed the integration of Automatic Speech Recognition, text summarization, and Text-to-Speech technologies for generating concise summaries from podcast audio. The research identified multilingual accessibility and low-resource language support as important challenges in Indian podcast summarization systems.

Another research work by Lokhande et al. presented an implementation framework for multilingual podcast summarization using transformer-based architectures [2]. The system utilized Whisper ASR for speech transcription and BART-based models for summary generation. Experimental observa-

tions indicated that transformer-based methods provide better contextual understanding and more meaningful summaries than conventional statistical approaches.

B. Automatic Speech Recognition Systems

Automatic Speech Recognition systems are essential components of podcast summarization because spoken audio must first be converted into text before further processing. Radford et al. introduced Whisper, a multilingual speech recognition model trained on large-scale audio datasets [3]. The model demonstrated strong transcription performance for multilingual speech and showed improved robustness in noisy audio environments. Whisper also supports several Indian languages, making it suitable for regional language podcast processing tasks.

Google Speech-to-Text API is another widely used speech recognition service supporting multilingual transcription [7]. Features such as automatic punctuation, speaker diarization, and real-time speech processing make it useful for audio analysis applications. However, dependence on internet connectivity and cloud infrastructure remains a limitation in low-resource environments.

C. Transformer-Based Summarization Models

Transformer architectures have significantly improved the quality of text summarization systems. Devlin et al. introduced BERT, a bidirectional transformer model capable of understanding contextual relationships between words and sentences [4]. BERT is commonly applied in extractive summarization tasks because it effectively identifies important textual information from transcripts.

Lewis et al. proposed BART, a transformer-based sequence-to-sequence model designed for abstractive text summarization [5]. Unlike extractive approaches, BART generates more natu-ral and human-readable summaries by understanding semantic relationships within text. Because of its strong contextual representation capability, BART has become widely used in podcast and document summarization research.

Multilingual transformer models such as mT5 and mBART further improved cross-lingual summarization and translation tasks [8, 9]. These models support multiple languages and provide better performance for multilingual datasets involving Indian regional languages.

D. NLP Frameworks for Indian Languages

Processing Indian regional languages requires specialized NLP frameworks because of script diversity, dialect variations, and limited language resources. Kakwani et al. introduced IndicNLPSuite and IndicBERT to improve multilingual NLP support for Indian languages [6]. The framework provides tokenization, transliteration, normalization, and language representation support for multiple regional languages including Hindi and Kannada.

Research initiatives by AI4Bharat also contributed to the development of multilingual datasets and transformer-based models for Indian language processing [10]. Such frameworks improved the feasibility of building multilingual podcast summarization systems for low-resource Indian languages.

E. Research Gaps in Existing Systems

Although recent studies have shown promising improve-ments in multilingual speech processing and summarization, several limitations continue to exist. Most existing systems are still optimized mainly for English-language datasets and show lower accuracy for regional Indian languages. Challenges such as code-mixed speech, accent variations, limited annotated datasets, and preservation of cultural context continue to affect summarization quality [1, 2].

Another major limitation is the computational complex-ity of transformer-based models. Large-scale models require significant processing power and memory resources, which makes deployment difficult in low-resource environments. In addition, comparatively fewer studies focus on Kannada and Telugu podcast summarization when compared to English and Hindi datasets. These limitations indicate the need for more regionally adaptive and computationally efficient multilingual podcast summarization systems.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 05 | May – 2026 DOI:10.55041/ISJEM07344
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

TABLE I
COMPARISON OF EXISTING RESEARCH WORKS

Ref.	Author(s) & Year	Dataset / Source	Languages	Methods / Models	Key Results	Limitations
[1]	Lokhande et al., 2024	Regional Podcasts	Hindi, Kannada, Telugu, Marathi, Tamil	Whisper + BART + TTS	ROUGE-L F1: 0.64	Limited dataset size
[2]	Lokhande et al., 2025	YouTube Podcasts	Hindi, Kannada, Telugu	Whisper + BART	Higher ROUGE scores	High computational cost
[3]	Radford et al., 2023	Speech Dataset	90+ languages	Whisper ASR	Robust multilingual ASR	Weak for code-mixed speech
[4]	Devlin et al., 2019	BooksCorpus + Wiki	English	BERT	Effective extractive summaries	Not abstractive
[5]	Lewis et al., 2020	CNN/DailyMail	English	BART	Fluent abstractive summaries	Needs high GPU
[6]	Kakwani et al., 2020	IndicCorp	Indian Languages	IndicBERT	Better Indian NLP support	Limited integration
[7]	Raffel et al., 2020	mC4 Corpus	101 languages	mT5	Cross-lingual summarization	Computationally expensive
[8]	Liu et al., 2020	CC25 Dataset	25+ languages	mBART	Strong multilingual generation	Needs fine-tuning
[9]	Mihalcea & Tarau, 2004	DUC-2004	English	TextRank	Simple extractive method	Lower semantic understanding
[10]	Erkan & Radev, 2004	DUC-2004	English	LexRank	Good for long documents	Redundant summaries

TABLE II
PERFORMANCE COMPARISON OF SUMMARIZATION MODELS

Model	Type	Dataset	ROUGE-1	ROUGE-2	ROUGE-L
BERT	Extractive	CNN/DailyMail	0.42	0.19	0.39
BART	Abstractive	CNN/DailyMail	0.45	0.21	0.41
mT5	Abstractive	XSum	0.44	0.20	0.40
mBART	Abstractive	XSum	0.46	0.22	0.42
TextRank	Extractive	DUC-2004	0.38	0.16	0.35
LexRank	Extractive	DUC-2004	0.40	0.17	0.35

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 05 | May – 2026 DOI:10.55041/ISJEM07344
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

III. METHODOLOGY

A. Podcast Audio Collection

The methodology begins with collecting podcast audio from platforms such as Spotify, YouTube Podcasts, and regional language streaming services. The collected podcasts may contain multilingual conversations in English, Kannada, Hindi, and Telugu. Since podcast recordings often include background noise, different speaking styles, and dialect variations, audio preprocessing is necessary before transcription [1, 2].

B. Audio Preprocessing

Audio preprocessing improves the quality of podcast recordings before speech recognition. Noise reduction, silence removal, audio normalization, and speech enhancement techniques are applied to improve transcription accuracy. Preprocessing is particularly important for multilingual podcasts because regional language audio may contain pronunciation differences and informal conversational patterns.

C. Automatic Speech Recognition

Automatic Speech Recognition (ASR) converts spoken pod-cast audio into textual transcripts. Whisper-based multilingual ASR models are commonly used because they support mul-tiple Indian languages and perform effectively for long-form audio transcription [3]. Speech recognition systems generate machine-readable transcripts that can later be processed using Natural Language Processing techniques.

D. Text Preprocessing

The generated transcripts undergo several preprocessing operations before summarization. NLP techniques such as tokenization, sentence segmentation, stop-word removal, and text normalization are applied to improve textual quality [6]. These operations help remove irrelevant information and improve contextual understanding within multilingual transcripts.

E. Abstractive Summarization

Abstractive summarization models generate concise summaries by understanding semantic relationships within textual content. Transformer-based models such as BART and mT5 are widely used for multilingual summarization because they generate more natural and context-aware summaries [5, 8]. These models improve readability while preserving the important meaning of podcast conversations.

F. Multilingual Translation

Machine translation techniques help convert summarized content into multiple regional languages based on user preference. Translation systems improve accessibility for users who prefer consuming summarized information in Kannada, Hindi, Telugu, or English. Multilingual transformer models further improve translation quality for low-resource Indian languages [6, 9].

G. Text-to-Speech Conversion

Text-to-Speech systems convert summarized text into au-dio format for voice-based accessibility. Speech synthesis improves usability for visually impaired users and supports multilingual audio playback. TTS systems also help users consume summarized information more conveniently in audio form.

H. Overall Workflow

The overall methodology combines multilingual ASR, NLP preprocessing, transformer-based summarization, multilingual translation, and speech synthesis techniques to generate con-cise summaries from podcast audio. The workflow improves multilingual accessibility and reduces the time required to consume long podcast episodes while preserving contextual meaning and regional language understanding.

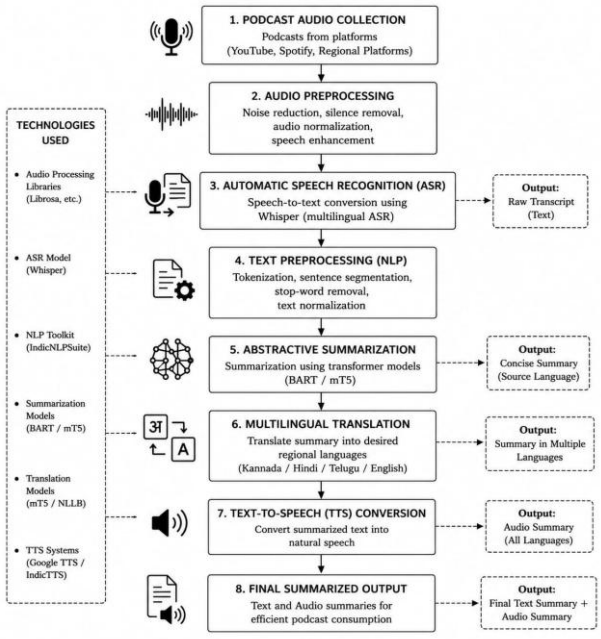

Figure: Overall Methodology for Multilingual Podcast Summarization

Fig. 3. Overall Methodology for Multilingual Podcast Summarization

IV. FUTURE SCOPE

Multilingual podcast summarization continues to be an emerging research area with significant scope for improvement in speech processing, language understanding, and regional language accessibility. Rapid advancements in Artificial Intel-ligence, Natural Language Processing, and transformer-based architectures are expected to improve the quality and efficiency of multilingual podcast summarization systems in the coming years.

One important research direction involves improving sup-port for low-resource Indian languages such as Kannada and Telugu. Existing summarization systems generally provide better performance for English-language datasets, while re-gional language processing still faces challenges due to limited annotated datasets and insufficient language resources. The

development of larger multilingual datasets and region-specific speech corpora may improve transcription and summarization accuracy for Indian languages [1, 2].

Another major area of future research includes improving code-mixed speech processing. Indian podcast conversations frequently involve language switching between English and regional languages within the same sentence. Advanced multilingual ASR models and transformer-based language models may help improve contextual understanding and reduce transcription errors in multilingual environments [3].

Future podcast summarization systems may also focus on real-time summarization of live audio streams. Real-time processing can improve applications related to news broadcasting, online education, multilingual meetings, and digital media platforms. Efficient lightweight transformer architectures may further support deployment in mobile and low-resource devices.

Emotion-aware summarization represents another promising research direction. Current summarization systems mainly focus on textual information while often ignoring emotional tone, speaker intent, and conversational emphasis. Integrating sentiment analysis and speech emotion recognition may improve the contextual quality and readability of generated summaries.

Multimodal summarization techniques involving audio, text, and video analysis may further improve summarization quality for multimedia podcast platforms. Combining speech processing with visual information extraction can help generate more informative and context-aware summaries for digital content applications.

Future advancements in multilingual translation systems are also expected to improve accessibility for regional language users. Improved translation models may help preserve contextual meaning, cultural expressions, and semantic relationships while generating summaries in multiple Indian languages [6]. Research related to Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) may further improve contextual summarization and multilingual understanding. Such technologies have the potential to generate more coherent, human-readable, and personalized summaries for long-duration podcast content.

The growing demand for multilingual digital communica-tion platforms indicates strong future potential for podcast summarization systems capable of supporting diverse regional language communities efficiently and accurately.

V. Conclusion

Multilingual podcast summarization has gained significant importance because of the rapid growth of regional language digital content and podcast platforms [1, 2]. The use of Artificial Intelligence, Automatic Speech Recognition, Natural Language Processing, and transformer-based language mod-els has improved the capability of processing multilingual podcast audio efficiently [3, 4]. Technologies such as Whis-per, Bidirectional Encoder Representations from Transformers, Bidirectional and Auto-Regressive Transformers, Multilingual Text-to-Text Transfer Transformer, and Indic Bidirectional Encoder Representations from Transformers have contributed to improved multilingual transcription, contextual understand-ing, and summary generation for Indian regional languages including Kannada, Hindi, Telugu, and English [3, 5, 6].

Existing research studies indicate that multilingual pod-cast summarization can improve information accessibility and reduce the time required to consume long-duration audio content [1, 2]. However, challenges related to code-mixed speech, regional dialect variations, pronunciation differences, and limited low-resource language datasets continue to affect summarization quality and multilingual language understand-ing [6].

Future advancements in multilingual Automatic Speech Recognition systems, transformer-based architectures, and regional language Natural Language Processing frameworks may further improve the efficiency, accessibility, and con-textual understanding of podcast summarization systems for Indian regional languages.

Acknowledgment

The authors acknowledge the support of the Department of Computer Science and Engineering, K. S. Institute of Technology, Bengaluru. We express our sincere gratitude to our project mentor, Dr. Vijayalaxmi Mekali, Professor, Department of Computer Science and Engineering, for her invaluable guidance and expertise throughout this work. The authors also thank the open-source communities behind Mozilla Common Voice, OpenAI Whisper, and AI4Bharat for making multilin-gual NLP research accessible.

References

[1] M. Lokhande et al. Comprehensive survey on podcast summarization in regional language. *ICSCNA*, 2024.
[2] M. Lokhande et al. Implementation of podcast summarization in regional languages. *WCONF*, 2025.
[3] A. Radford et al. Robust speech recognition via large-scale weak supervision. *OpenAI Whisper Research*, 2023.
[4] J. Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.
[5] M. Lewis et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation. *ACL*, 2020.
[6] D. Kakwani et al. Indicnlpsuite: Multilingual language models for indian languages. *EMNLP Findings*, 2020.
[7] Google Cloud. Google speech-to-text api documentation, 2024.
[8] C. Raffel et al. Exploring the limits of transfer learn-ing with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.
[9] Hugging Face. mbart multilingual sequence-to-sequence model, 2024.
[10] AI4Bharat. Indicbert and indian language resources, 2024.