International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 05 | May – 2026 DOI:10.55041/ISJEM07358
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Aawaz Aur Drishya Ka Setu: Podcast Audio-To-Image Generation using Automatic Speech Recognition, Summarization, and Generative AI

Dr. Vijayalaxmi Mekali, Akash S, Amruth CK, C Nagendra Reddy, G Akash
Department of Computer Science and Engineering
K. S. Institute of Technology, Bengaluru – 560109, India vijayalaxmimekali@ksit.edu.in
{1KS23CS007, 1KS23CS010, 1KS23CS032, 1KS23CS044}@ksit.edu.in

Abstract- Podcasts have become one of the most popular platforms for sharing knowledge, storytelling, education, and entertainment. However, podcasts mainly rely on audio, making it difficult for users to visually understand key topics and follow long conversations. This project, "Aawaz aur Drishya ka Setu", introduces an AI-based system that transforms podcast audio into meaningful visual representations. The system converts podcast speech into text using speech recognition techniques and divides the transcript into meaningful sections. Natural Language Processing is then used to summarize content, extract keywords, and identify emotional tones. Based on the detected context and emotions, AI image generation models create visuals that represent the podcast scenes and mood. Color theory is also applied to improve emotional connection and storytelling. The proposed system enhances podcast accessibility, user engagement, and content presentation for creators, educational platforms, media applications, and hearing-impaired users.

Keywords — Podcast Visualization, Speech Recognition, Natural Language processing, Image Generation, Emotion Detection, Multimodal Learning.

INTRODUCTION

Podcasts have become a popular digital medium for communication, learning, interviews, entertainment, and storytelling. Millions listen to podcasts every day because they are easy to access while traveling, working, or doing daily tasks. However, podcasts rely heavily on audio, which creates limitations for users. Long episodes are hard to navigate, key moments are hard to spot, and users often struggle to quickly grasp the content without listening to the entire audio.

In today's digital landscape, visual content captures attention more effectively than audio alone. Social media platforms, educational apps, and content-sharing websites mainly use images, thumbnails, and short visual summaries to boost engagement. Traditional podcasts lack this visual support, making it harder for creators to promote content and for users to find topics of interest.(Kurniawan et al., 2024) Accessibility is another major challenge. Hearing-impaired users cannot fully enjoy podcast content because it is only available in audio form. Current podcast systems mainly focus on transcribing speech to text or basic summarization, but they do not provide meaningful visual representations based on the podcast's emotional and contextual content.

To tackle these issues, this project proposes "Aawaz aur Drishya ka Setu," which translates to "Bridge Between Sound and Vision." The project's main goal is to automatically create visuals from podcast audio using Artificial Intelligence methods. (Xiao et al., 2025) The system changes speech into text, examines the content's emotional and semantic meaning, and generates images that represent key scenes, moods, and concepts discussed in the podcast.

The proposed system also includes chunk-based processing, where long podcasts are divided into smaller sections based on topic flow and emotional changes. This helps maintain context and improves visual accuracy. The project also applies color theory principles during image generation to enrich emotional storytelling and user engagement.(Jin et al., 2025) The system can help podcast creators save time on editing while reaching a wider audience on social media. It can also enhance accessibility and make long audio content easier to understand through visual storytelling.

LITERATURE SURVEY

Recent research in Artificial Intelligence and multimedia processing has shown significantprogress in podcast understanding, speech recognition, summarization, and image creation. Many researchers focus on improving podcast accessibility and content navigation through AI-driven methods.

Jimin Park et al. proposed a system to improve the podcast browsing experience through topic segmentation and visualization using Generative AI.(Park et al., 2024) Their work concentrated on breaking podcasts into meaningful sections and creating visual cues for easier navigation. The study revealed that users often struggle to understand long podcast content using just metadata and transcripts. The addition of visual elements improved browsing efficiency and user engagement. However, their system mainly focused on navigation and did not explore emotional understanding or visual storytelling deeply. Another key research area is Automatic Speech Recognition (ASR). OpenAI Whisper models have shown strong capabilities in recognizing multilingual speech and processing noisy audio environments. Whisper can reliably convert podcast audio into text, handling different accents, speaking speeds, and conversational styles. This makes it a suitable choice for long-form podcast applications.(Liang et al., 2024)

Research on transformer-based summarization models like Bidirectional and Auto-Regressive Transformers has significantly improved Natural Language Processing tasks. Such models are highly effective in generating concise summaries from lengthy transcripts by removing redundant information. However, lengthy podcast transcripts often surpass the token capacity of standard transformer architectures, creating difficulties in processing and maintaining contextual continuity.

Harini Narayanan proposed a system for assessing creativity in AI-generated podcasts. The research introduced concepts like emotional diversity, analogy detection, and divergent thinking analysis in podcast content(Narayanan, 2025).

It illustrated how emotional understanding can enhance the quality and creativity of AI-generated media. Yet, this research focused more on creativity measurement than on visual interpretations of podcast content.

Recent advances in text-to-image generation models like Stable Diffusion and DALL-E now allow for high-quality image creation from text prompts. These models can produce visually rich outputs based on semantic meaning. Many current systems apply image generation only for simple prompt-based outputs and lack the ability to maintain a continuous narrative across long audio content.

Most existing systems handle individual tasks, such as transcription, summarization, or image generation, separately. Few combine speech recognition, emotional analysis, semantic understanding, and AI-generated visualization into a single process.This gap highlights the need for systems that can turn long podcast audio into emotionally resonant visual storytelling experiences.

PROPOSED SYSTEM

The proposed work, “Aawaz aur Drishya ka Setu,” aims to turn podcast audio into meaningful visual representations using Artificial Intelligence techniques. The system links audio content and visual storytelling by integrating speech recognition, Natural Language Processing, emotional analysis, and AI image generation.

The process begins with podcast audio, YouTube podcast links, or uploaded video files as input. The system extracts audio from the video using tools like FFmpeg and yt-dlp. Once the extraction is complete, the audio is processed with speech recognition models to create text transcripts from the spoken content.(Hossain et al., 2025)

To enhance understanding of long-form content, the transcript is divided into smaller meaningful chunks based on topic shifts, pauses, emotional changes, and narration flow. This chunk-based processing helps maintain contextual continuity and improves output accuracy.

After segmentation, the system applies summarization and keyword extraction techniques using NLP models. Key topics, keywords, and meanings are identified from each chunk. Along with semantic analysis, the system also detects emotions by analyzing speech patterns, voice pitch, pauses, and tonal variations. Emotions such as happiness, sadness, suspense, confidence, and motivation are recognized to enhance storytelling quality.

Using the extracted context and emotions, prompts are generated for the image generation model. AI- based image generation systems create visuals that reflect the podcast's content, mood, and narrative development.(Park et al., 2024) The images produced are not just literal representations but also emotionally aligned interpretations.

The proposed work incorporates color theory during image generation. Different color palettes are applied based on emotional tone to improve audience engagement and emotional connection. For example, motivational segments might feature warm colors, while suspenseful scenes could use darker tones and shadows. Ultimately, the generated visuals, summaries, and transcripts are displayed through a user interface that

allows users to explore podcasts visually. This system helps users quickly understand podcast flow while improving accessibility for those with hearing impairments.

METHODOLOGY

The methodology of the system involves several interconnected stages responsible for transforming podcast audio content into visual storytelling formats.

A. Audio Collection and Preprocessing

The system is designed to take podcast audio inputs from YouTube links, uploaded audio files, or video files. It uses FFmpeg to extract audio from video, converting it to a suitable format, and ensuring appropriate sampling rates and noise reduction.

B. Speech-to-Text Conversion

The cleaned audio is then fed into the Whisper speech recognition model, which accurately transcribes conversations, even in noisy environments and multilingual settings(Wu et al., 2024). Timestamp data is retained for synchronizing audio with visual elements.

C. Chunk-Based Segmentation

Long podcast transcripts are segmented into smaller meaningful chunks based on factors like pauses, changes in topic, shifts in emotion, and the narration flow. This aids the system in maintaining context and processing extended audio efficiently.

D. Summarization and Keyword Extraction Each chunk's content is then summarized into a concise representation using NLP techniques. Models like BART and T5 are employed for effective summarization, extracting key keywords, named entities, and themes.

E. Emotion Detection

The system analyzes voice pitch, speech pace, tonal variations, and language context to detect emotions like excitement, sadness, motivation, fear, suspense, and curiosity within each chunk, ensuring relevance and enhanced storytelling in generated images.

F. Prompt Generation

The summaries, keywords, and detected emotions are used to automatically generate prompts for the image generation model, encompassing elements like themes, settings, mood, and artistic style.(Xue et al., 2024)

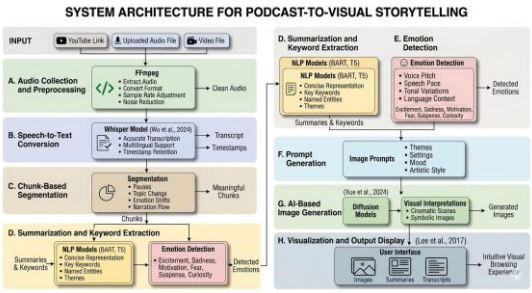
G. AI-Based Image Generation

These prompts are then fed into diffusion-based image generation models to produce various visual interpretations of podcast segments, ranging from cinematic scenes to symbolic images.

H. Visualization and Output Display

The final outputs (images, summaries, transcripts) are presented through a userinterface, creating an intuitive visual browsing experience for podcast content. (Lee et al., 2017)

SYSTEM ARCHITECTURE:



IMPLEMENTATION

The implementation of the system is carried out using Python along with several AI libraries. The system is structured as a modular pipeline, where each module handles a specific aspect of podcast processing and visualization.

The frontend of the system is built using Streamlit to offer an interactive and user-friendly interface where users can upload files or paste YouTube links to podcasts. The backend is responsible for executing the audio extraction, transcription, summarization, emotion detection, and image generation processes.

For audio extraction from videos and online podcasts, FFmpeg and yt-dlp are utilized. OpenAI Whisper is chosen for its accurate and robust speech-to-text conversion capabilities, (Hossain et al., 2025)especially for conversational audio and multilingual support.

To perform text processing and summarization, transformer-based NLP models such as BART and T5 are used. Keyword extraction and semantic analysis are performed using NLP libraries like spaCy and NLTK. Speech analysis and contextual semantics are employed for emotion detection.(Udawant et al., 2025) For the image generation component,diffusion models, specifically Stable Diffusion, are leveraged to translate textual prompts into high-fidelity visual representations (Kumar, 2025; Nambiar, 2024).

The summaries and detected emotions are transformed into prompts that are then processed by diffusion-based image generation models to create corresponding visual outputs.

The final system presents transcripts, summaries, generated images, and topic-wise visualizations in a dashboard. The chunk-wise processing enables the system to efficiently manage long podcast episodes.

RESULTS AND DISCUSSION

The system successfully transformed podcast audio into coherent visual representations, demonstrating the potential of AI for enhancing podcast accessibility, engagement, and comprehension.

The speech recognition module produced accurate transcripts under varying audio conditions. The chunk-based segmentation improved contextual understanding and facilitated efficient processing of long-form content.

The summarization module effectively condensed lengthy transcripts into concise and informative summaries. The emotion detection feature enriched the visual outputs by aligning them with the emotional tone of the podcast segments. Generated images depicted themes such as motivation, suspense, storytelling, and the overall emotional trajectory of the podcast.

The integration of color theory enhanced the emotional impact of the visuals. Users gained a better grasp of the podcast's mood and narrative progression through visual storytelling compared to purely text-based transcripts.

The system also automated content creation tasks for podcast creators, producing visuals suitable for thumbnails, storyboards, and social media. It also improved accessibility for hearing-impaired individuals by enabling visual understanding of podcast content.

Some limitations were encountered during the implementation, including the high computational resources required for long podcasts, variations in image generation quality based on prompt precision and model capabilities, and potential inaccuracies in emotion detection due to speaker tone or background noise.

Overall, this project highlighted the ability to combine speech recognition, NLP, emotional analysis, and AI-generated visualization to create an engaging and multimodal podcast experience.

Future Scope

Future improvements to the system could include real-time podcast visualization, support for multiple languages, and the generation of animated videos instead of static images to offer a more immersive storytelling experience.

More sophisticated emotion detection models that can identify nuances like sarcasm and humor, as well as deep emotional transitions, could further enhance the system's understanding. Integration with AR and VR platforms could lead to interactive visual podcast experiences.

The system could also be adapted for educational purposes, audiobook visualization, news summarization, and AI-assisted storytelling platforms.

CONCLUSION

The presented paper introduced "Aawaz aur Drishya ka Setu," an AI-based system designed to convert podcast audio into meaningful visual representations. The system seamlessly integrates speech recognition, Natural Language Processing, emotion analysis, and generative AI to transform long-form podcast content into a multimodal storytelling experience.

The system effectively performs speech-to-text conversion, chunk-based segmentation, summarization, emotion detection, and image generation, resulting in visually clear podcast outputs. The incorporation of color theory and emotion analysis further elevated storytelling quality and audience engagement.

The system aims to improve podcast understanding and accessibility for hearing-impaired users, while also reducing manual effort for creators and enhancing content discoverability.

Despite some limitations in computational complexity and image generation consistency, the project demonstrates significant potential for future multimedia applications. Potential enhancements include real-time processing, multilingual visual generation, advanced emotional intelligence, and animated video storytelling.

This project underscores the power of Artificial Intelligence in bridging sound and vision by transforming audio content into immersive visual experiences.

REFERENCES

[1] Jimin Park, Chaerin Lee, Eunbin Cho, and Uran Oh, "Enhancing the Podcast Browsing Experience through Topic Segmentation and Visualization with Generative AI," ACM International Conference on Interactive Media Experiences (IMX '24), 2024.

[2] Harini Narayanan, "Quantifying Creativity in AI-Generated Podcasts," IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC), 2025.

[3] Alec Radford et al., "Robust Speech Recognition via Large- Scale Weak Supervision," OpenAI Whisper Research Paper, 2022.

[4] Mike Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation," ACL, 2020.

[5] Colin Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research, 2020.

[6] Robin Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models," CVPR, 2022.

[7] Tom Brown et al., "Language Models are Few-Shot Learners," NeurIPS, 2020.