Journal of Science and Technology
ISSN: 2456-5660 Volume 10, Issue 01 (Jan -2025)
www.jst.org.in *DOI:https://doi.org/10.46243/jst.2025.v10.i01.pp10- 25*

Revolutionizing Audio Content Navigation: AI-Enhanced Multimodality and Machine Learning for Speaker Diarization and Topic Segmentation

Waseem Syed

Senior Staff Software Engineer,
Intuit, CA, United States

To Cite this Article

Waseem Syed**"Revolutionizing Audio Content Navigation: AI-Enhanced Multimodality and Machine Learning for Speaker Diarization and Topic Segmentation**''. *Journal of Science and Technology, Vol. 10, Issue 01-Jan 2025, pp01-09*

Article Info

Received: 30-10-2024 Revised: 22-12-2024 Accepted: 10-01-2025 Published:25-01-2025

Abstract. The rapid evolution of digital media has propelled an increased consumption of audio content, ranging from podcasts to educational lectures. Despite its growing popularity, the inherent unstructured nature of audio media poses significant challenges in navigation and user interaction. Our paper introduces an innovative AI-driven framework designed to fundamentally transform audio content exploration. Utilizing cutting-edge machine learning and deep learning technologies, the system applies precise speaker diarization and topic segmentation to radically improve navigation and content discovery in audio streams. Furthermore, the incorporation of an interactive chat feature enriches user interaction, allowing listeners to effortlessly query and jump directly to specific content via intuitive voice and text commands. This advanced system not only streamlines the audio exploration process but also personalizes the listener experience by integrating multimodal interfaces and sophisticated content annotation techniques. By addressing critical navigational inefficiencies, this framework sets a new paradigm in personalized, structured, and interactive media consumption, catering to the evolving demands of modern audio content users.

Keywords: Speaker Diarization, Topic Segmentation, Artificial Intelligence(AI), Machine Learning, Multimodal Processing, Natural Language Processing(NLP), Content Navigation, Deep Learning.

1 Introduction

The rapid growth of digital audio platforms, like podcasts, has revealed significant navigation challenges due to their unstructured format. This paper introduces a targeted framework powered by advanced machine learning technologies, such as speaker diarization [1,2] and AI-driven contextual topic modeling [3,4]. Enhanced by OpenAI Whisper and Google Gemini, our approach improves speaker detection in multi-speaker settings, offering personalized, dynamic navigation aids tailored to user preferences [5,8,13]. Despite the widespread appeal of podcasts for delivering long-form content with on-demand access, conventional navigation tools underperform, failing to adequately serve listeners who prefer specific segments over full episodes [10,11,12]. By integrating audio, text, and visual elements, our framework overcomes these traditional shortcomings, enabling a more interactive listening experience essential for users focused on specific content parts [14, 15]. This introduction encapsulates the development, functionality, and potential broader impacts of these innovations on multimedia interaction.

2 Literature Review

2.1 User Behavior in Audio Consumption

Research indicates a growing preference among listeners for targeted access to audio segments prioritizing specific content over full episodes [16]. Speaker diarization significantly enhances this by accurately identifying speaker turns in multi-speaker settings [2,17].

2.2 Advances in AI for Audio Content Analysis

Deep learning has markedly improved speaker diarization, topic segmentation, and content annotation [18,19]. Notably, new diarization models, including the system developed by the BUT team for the VoxCeleb Speaker Recognition Challenge, have achieved remarkably low Diarization Error Rates, with the BUT team system reaching a DER of 4% on the VoxConverse dataset [20], while advanced neural models

for topic segmentation now achieve high accuracy [21]. Additionally, the Deep Neural Network-based system reached an accuracy of 96.24%, precision for true positive detection (correct speech detections) of 0.978, and recall of 0.981 signifying a significantly high speech detection where there is indeed speech in the audio frame[20].

2.3 Generative AI in Multi-Modal Interfaces

Integrating audio, text, and visuals benefits user interaction [22]. Combining speech recognition with text processing enhances scalability [19], and collaborations between humans and AI generate enriched multimedia content summaries, proving essential for enhanced engagement. Generative AI, including models like GPT-4, plays a crucial role in dynamic podcast indexing and tailored content creation, such as generating customized teaser clips for improved content. Implementing generative AI for voice-to-text queries also elevates accessibility and user interaction [23,24].

2.4 Current Audio Navigation Interfaces

Current interfaces offer basic functionalities like timestamps but lack advanced personalization. The application of generative AI for creating contextual annotations could revolutionize these interfaces, with added features like ratings and analytics enhancing user experience management [7,8].

3 High-Level Approach

Our approach aims to categorize and index audio content effectively, providing listeners with targeted access to desired segments. This structured interaction transforms passive listening into active exploration of content, catering to user demands for immediacy and relevance. The approach consists of several techniques:

3.1 Speaker Diarization

Displays a list of all speakers featured in the podcast episode, extracted directly from the audio using advanced speaker diarization.
Features:

Journal of Science and Technology
ISSN: 2456-5660 Volume 10, Issue 01 (Jan -2025)
www.jst.org.in *DOI:https://doi.org/10.46243/jst.2025.v10.i01.pp10- 25*

- **Dynamic Speaker Profiles:** Clicking on a speaker's name opens a mini profile displaying a photo, brief biography, and notable contributions during the episode.
- **Speaker Popularity Metrics:** Options to show metrics like the number of times each speaker was engaged in discussions.

Fig. 1. Speaker Diarization

Speaker diarization determines "who spoke when" by minimizing errors in speaker identification using the Diarization Error Rate (DER) ensuring accurate breakdown of speaker contributions.

$$\text{DER} = \frac{FA + Miss + Conf}{Total\ Time}$$

- FA (False Alarm): Time wrongly attributed to a speaker.
- Miss: Time a speaker is not detected.
- Conf (Confusion): Overlapping time segments attributed to the wrong speaker.
- Total Time: Duration of the podcast.

3.2 Topic Segmentation

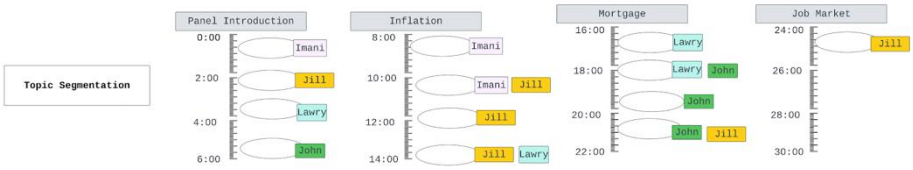
Displays a comprehensive list of topics discussed with timestamps using NLP techniques for high-accuracy topic segmentation.

Features:

- **Topic-Based Indexing:** Each topic is hyperlinked, allowing listeners to jump directly to the segment of interest.
- **Visual Indicators:** Visual cues (icons) indicate the nature of discussions, such as critical insights (star), personal anecdotes (heart), or technical data (graph).

Journal of Science and Technology
ISSN: 2456-5660 Volume 10, Issue 01 (Jan -2025)
www.jst.org.in *DOI:https://doi.org/10.46243/jst.2025.v10.i01.pp10- 25*

Fig. 2. Topic Segmentation



Topic segmentation leverages Cosine Similarity to determine the similarity between sentences or sections, enabling the grouping of related content under unified topics ensuring related sentences are clustered for precise topic categorization

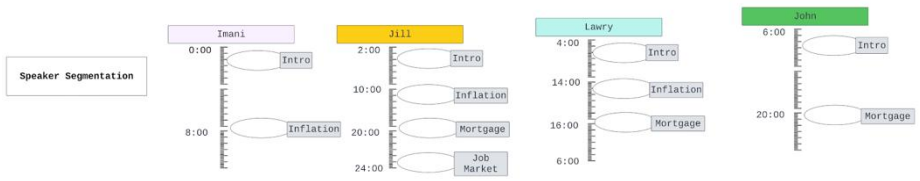
$$Consine\ Similarity = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A^2{}_i} \cdot \sqrt{\sum_{i=1}^{n} B^2{}_i}}$$

- A and B: Text vector representations of two sentences or sections.
- n : Number of features (e.g., keywords) in the vectors.

3.3 Speaker Segmentation

Ensures precise attribution of each speaker's contributions by aligning timestamps accurately, maintaining clear identification of speaker turns

Fig. 3. Speaker Segmentation



$$t_s = t_{start} + \triangle t$$

- t_s = Timestamp of speaker's segment
- t_{start} = Starting time of audio
- $\triangle t$ = Offset calculated using diarization models

3.4 Multi-Modal Search Interface

This interface supports user queries through text and voice inputs for streamlined search functionality.

Features:

- **Interactive Chat**: Users type queries such as, "Did Tom speak about the housing market?" to retrieve relevant audio segments.
- **Voice Processing**: Spoken queries are transcribed into text using Apple's Speech Framework for iOS or Google's Speech API for Android.
- **AI Integration**: For text inquiries, OpenAI's GPT model (such as GPT-4) processes the queries via an API call to ChatGPT, generating a list of pertinent segments with minimal latency.

3.5 Advanced Playback

This layer enhances playback by integrating contextual data into selected segments.

Featues:

- **Dynamic Annotations**: Annotations appear during playback, highlighting key points, references, or statistics.
- **Follow-Up Links**: Offers links to related episodes or relevant articles based on current topics.
- **Integrated Notes**: Allows users to take timestamped notes during playback for future reference.

The relevance of topics in playback is determined using:

$$Relevance\ Score = TF - IDF(k) \times Weight(k)$$

- TF-IDF (Term Frequency-Inverse Document Frequency): Calculates the importance of a keyword k in the transcript.
- Weight(k): Adjusts the relevance based on speaker emphasis or repetitions.

3.6 Feedback and Engagement

This layer enhances listener engagement by incorporating user feedback mechanisms.

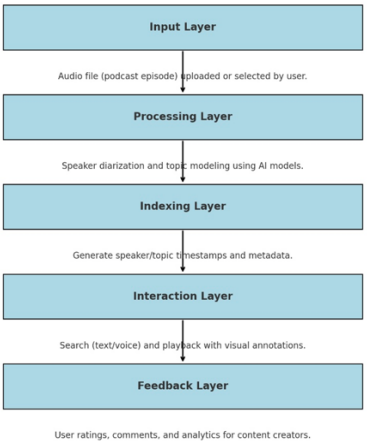
Features:

- **Rating System:** After engaging with a segment or speaker, users can rate their experience on a scale (e.g., 1 to 5 stars).
- **Comment Section:** Allow users to share thoughts or questions in a threaded discussion format for each topic or segment.
- **Analytics Dashboard:** Content creators can analyze feedback trends, helping them understand listener preferences and improving future content.

4 Implementation

4.1 Service Layer

Fig. 4. System Architecture



4.1.1 Input Layer

The Input Layer accepts uploaded audio files, which trigger the initial processing steps necessary for speaker diarization and topic modeling using OpenAI's Whisper. In our observations, Whisper outperforms

other models in terms of accuracy of timestamps and synchronization with the correct playback speed.

Table 1. Algorithm Transcribe Audio

Stage	**Operation**
Input	*audio_file_path* (string)
Output	JSON response with transcription data or error
Procedure	**transcribe(audio_file_path):**
1. Try Block	– Open audio_file in binary mode. – Call API with: file, 'whisper-1', 'verbose_json', ['word']. – Store response in transcript.
2. Catch	Return JSON with "File not found" error and code 404.
3. Catch Other Exceptions	Return JSON with error detail and code 500.
4. Extract & Compile	– Extract details from transcript. – Compile and return JSON with transcription_data.
End Procedure	

Table 2. Sample Response – Transcribe

{
"duration": 25.02,
"language": "english",
"task": "transcribe",
"text": "All right, so hi everyone, my name is John Doe, and I'm a Senior Business Analyst at ... Let's move on to Jane...
... Thanks, John. Hello everyone, my name is Jane Doe, and I'm a Principal Advisor ..."
"words": [
{"start": 1.96, "end": 2.56, "word": "All"},
{"start": 2.56, "end": 2.74, "word": "right"},
{"start": 3.04, "end": 3.36, "word": "so"}, ...
]
}

4.1.2 Processing Layer

This layer leverages advanced AI for precise Speaker Diarization and Topic Segmentation. Diarization identifies speakers and their timing, assessed by Diarization Error Rate (DER). Topic segmentation uses cosine similarity to cluster content by topics efficiently.

Table 3. Algorithm Categorize Speakers and Topics implementation

Stage	Operation
Input	*transcription_data* (JSON object with text transcriptions and timestamps)
Output	JSON with categorized speaker/topic data or error message with status code
Procedure	**analyze_transcription**(transcription_data):
1. Formulate Prompt	Create prompt from **transcription_data** for speaker/topic inference.
2. Try Block	─ Invoke GPT-4 with model "gpt-4" and formulated prompt. ─ Convert GPT-4 output into JSON and extract structured data.
3. Catch	Return JSON with "Failed to analyze data" and code 500.
End Procedure	

Table 4. Sample Response – Speaker/Topic Segmentation

```
{
  "speakers": [
    {
      "name": "John Doe",
      "utterances": [
        {
          "text": "All right,so hi..",
          "start": 1.96,
          "end": 14.04
        },
        {
          "text": "Global Economy...",
          "start": 14.20,
          "end": 20.15
        }
      ]
    },
    {
      "name": "Jane Doe",
      "utterances": [
```

```
            {
              "text": "Thanks, John...",
                "start": 15.70,
                "end": 21.42
            },
            {
              "text": "The job market.. ",
                "start": 21.50,
                "end": 27.30
            }
          ]
      }
... ...
   ],
   "topics": [
      {
       "name": "Global Economy",
       "utterances": [
          {
           "speaker": "John Doe",
           "text": "The global economy.. ",
               "start": 14.20,
               "end": 20.15
           }
          ]
      },
      {
       "name": "Job Market",
       "utterances": [
         {
            "speaker": "Jane Doe",
            "text": "The job market is...",
               "start": 21.50,
               "end": 27.30
            }
          ]
      }
... ...
   ]
```

| }
|---|

4.1.3 Indexing Layer

Metadata extracted from the previous layers (such as speaker labels, timestamps, and topics) is structured into a searchable index, facilitating efficient query processing and retrieval.

4.1.4 Interaction Layer

This user-facing layer allows interaction via text and voice inputs, using a multimodal search interface for immediate navigation of indexed podcast content. Users can issue queries in text or audio, with audio inputs converted to text through speech recognition software like Apple's Speech Framework or Google's Speech API.

Table 5. Algorithm – Retrieve Query Based Segments

Stage	**Operation**
Input	*transcription_data* (JSON object with text transcriptions), *query*
Output	JSON with relevant transcription segments
Procedure	**retrieve_segments(transcription_data, query):**
1. Execute GPT 4(or other AI model)	– Utilize the GPT-4 chat model process with a system message instructing to "Retrieve segments from transcript based on query." – User message is formatted by combining the search query with transcription data.
2. Return Data	– Parse and return JSON structured segments derived from the model's output – Return empty content if the model fails to retrieve relevant segments.
End Procedure	

Table 6. Sample Response – Query

{ "segments": [

```
    {
     "speaker_name": "Jane Doe",
     "sentence_spoken": "The job market..",
        "start_time": 21.50,
        "end_time": 27.30
    },
    {
     "speaker_name": "John
     "sentence_spoken": "Jane mentioned",
        "start_time": 27.35,
        "end_time": 29.00
     }
... ...
   ]
}
```


4.1.5 Feedback Layer

This layer collects user interactions and feedback through a rating system, commenting capabilities, and a comprehensive analytics dashboard that tracks engagement metrics. It allows users to rate segments, leave comments, and pose follow-up questions, enhancing engagement. The dashboard leverages this data to analyze user preferences and behaviors regarding different speakers and topics, guiding future content improvements.

4.2 UI Layer

Provide intuitive user interfaces for smooth interaction with backend services including:

- **Automatic Sliders:** Navigate directly to audio segments when topics or timestamps are selected. Example: Clicking "Global Economy by John Doe" moves playback to the corresponding discussion.
- **Contextual UI with Categorization:** Display organized speaker names and topics with visual elements for easy identification. Example: Icons next to names and topics help users quickly identify speakers and content themes.

- **Chat Input with Suggestions:** Allow quick query entry with auto-suggest for efficiency. Example: Typing "health" prompts suggestions like "Health Care by Jane Doe", facilitating faster access.
- **Voice Input:** Implement a microphone icon for inputting commands or searches via voice. Example: User says, "Show me where Jane discusses healthcare," and the system outputs relevant segments.
- **Social Media Sharing:** Enable direct sharing of snippets to social media, complemented by easy-to-use share icons. Example: Users can share a segment discussing "Job Market Trends" directly on their social media platforms.
- **Accessibility Support:** Enhances social inclusion, offers intuitive navigation aids.

5 Future Research

While the results for our approach are awaited, it is noteworthy that Edison Research indicates significant growth in the podcasting landscape, underscoring the increasing demand for advanced audio content navigation systems [25]

Table 6. Share of time spent listening to audio sources (US Population 13+)[25]

Audio Source	**Time Spent Listening (%)**
AM/FM	36
Sirius XM	8
TV Music Channels	3
Owned Music (CDs, DVDs, music files etc.)	7
Streaming Music (Spotify, Pandora, Apple Music, Amazon Music etc.)	18
Podcasts	10
YouTube Music/Music Videos	14
Audiobooks	3
Other	1

More than 40% of listeners utilized some form of online audio platform in 2023, marking a significant increase in users who require personalized audio navigation. This upward trend in podcast consumption highlights the necessity for innovative systems that enhance user experience. As the

audience expands, listeners seek more efficient ways to navigate and interact with content, reinforcing the relevance of AI-driven frameworks for speaker diarization and topic segmentation.

Future research avenues could explore the integration of AI-driven recommendation engines that suggest relevant segments based on user preferences and listening history. Additionally, expanding our approach to accommodate other audio formats, such as webinars and interviews, presents exciting opportunities to enhance content accessibility across diverse domains. However, challenges with these formats may include variations in audio quality and differences in discourse structures.

6 Conclusion

This paper presents a transformative framework that employs generative AI and advanced machine learning techniques to enhance the navigation and interaction of podcast content, setting a new standard in personalized media consumption. By transforming unstructured audio into modular, searchable units, the approach significantly enhances the user experience, making content exploration both intuitive and comprehensive. With sophisticated features like voice queries, smart speaker identification, and dynamic annotations, the framework not only improves accessibility and engagement but also provides valuable insights for content creators. Moving forward, its potential application across various audio formats promises further advancements in digital media accessibility, paving the way for more interactive and user-centric audio experiences. Future collaborative developments and research will be crucial in realizing the full potential of AI-driven audio content systems, marking a significant step forward in how we interact with and consume media.

7 References

1. Anguera X, Bozonnet S, Evans N, Fredouille C, Friedland G, Vinyals O. Speaker diarization: A review of recent research. IEEE Transactions on Audio, Speech, and Language Processing. 2012;20(2):356-370. doi:10.1109/TASL.2011.2125954.
2. Park TJ, Kanda N, Dimitriadis D, Han KJ, Watanabe S. A review of speaker diarization: Recent advances with deep learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021;29:1086-1103. doi:10.1109/TASLP.2021.3066309.

3. Mao HH, Li S, McAuley J, Cottrell G. Speech recognition and multi-speaker diarization of long conversations. arXiv preprint. 2020;arXiv:2005.08072.
4. Frølund, J., Løkke, A., Jensen, H., & Farver-Vestergaard, I. (2024). Development of Podcasts in a Hospital Setting: A User-Centered Approach. Journal of Health Communication, 29(4), 244–255. https://doi.org/10.1080/10810730.2024.2321385.
5. Vallet F, Essid S, Carrive J. A multimodal approach to speaker diarization on TV talk-shows. IEEE Transactions on Multimedia. 2013;15(3):509-520. doi:10.1109/TMM.2012.2233724.
6. Wang S, Ning Z, Truong A, et al. PodReels: Human-AI co-creation of video podcast teasers. Proceedings of the 2024 Designing Interactive Systems Conference (DIS '24). ACM. 2024;17 pages. doi:10.1145/3643834.3661591.
7. Austin A, Samuel A. Enhancing podcasting by leveraging AI technologies. Communications of the ACM. 2023;66(10):48-55. doi:10.1145/3626767.3625304.
8. Edison Research. Podcast Consumer Report. Edison Research. 2023.
9. Kinnunen T, Li H. An overview of text-independent speaker recognition: From features to supervectors. Speech Communication. 2010;52(1):12-40. doi:10.1016/j.specom.2009.08.009.
10. Chan-Olmsted, S., & Wang, R. (2022). Understanding podcast users: Consumption motives and behaviors. New Media & Society, 24(3), 684-704. https://doi.org/10.1177/1461444820963776.
11. Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Association for Computing Machinery, New York, NY, USA, Paper 640, 1–12. https://doi.org/10.1145/3173574.3174214.
12. Iacopo Ghinassi, Lin Wang, Chris Newell, and Matthew Purver. 2023. Multimodal Topic Segmentation of Podcast Shows with Pre-trained Neural Encoders. In Proceedings of the 2023 ACM International Conference on Multimedia Retrieval (ICMR '23). Association for Computing Machinery, New York, NY, USA, 602–606. https://doi.org/10.1145/3591106.3592270.
13. A. Hajavi and A. Etemad, "Fine-grained Early Frequency Attention for Deep Speaker Recognition," 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 2022, pp. 1-6, doi: 10.1109/IJCNN55064.2022.9892054.
14. U. S. Jha, "Efficient multimodal signal processing engine ease communication, computing, and multimedia convergence," 2005 IEEE International Conference on Personal Wireless Communications, 2005. ICPWC 2005., New Delhi, India, 2005, pp. 348-352, doi: 10.1109/ICPWC.2005.1431364
15. Galli, Carlo, et al. "Topic Modeling for Faster Literature Screening Using Transformer-Based Embeddings." Metrics. Vol. 1. No. 1. MDPI, 2024.
16. Lie Lu, Hong-Jiang Zhang and Hao Jiang, "Content analysis for audio classification and segmentation," in IEEE Transactions on Speech and Audio Processing, vol. 10, no. 7, pp. 504-516, Oct. 2002, doi: 10.1109/TSA.2002.804546.
17. A. Gomez, M. S. Pattichis and S. Celedón-Pattichis, "Speaker Diarization and Identification From Single Channel Classroom Audio Recordings Using Virtual Microphones," in IEEE Access, vol. 10, pp. 56256-56266, 2022, doi: 10.1109/ACCESS.2022.3177584..
18. Ghosh, R. et al. (2024). Topic Segmentation of Semi-structured and Unstructured Conversational Datasets Using Language Models. In: Arai, K. (eds) Intelligent Systems and Applications. IntelliSys 2023. Lecture Notes in Networks and Systems, vol 825. Springer, Cham. https://doi.org/10.1007/978-3-031-47718-8_7.
19. Deepika Yadav, Mayank Gupta, Malolan Chetlur, and Pushpendra Singh. 2018. Automatic Annotation of Voice Forum Content for Rural Users and Evaluation of Relevance. In

Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS '18). Association for Computing Machinery, New York, NY, USA, Article 12, 1–11. https://doi.org/10.1145/3209811.3209875.

20. F. Landini et al., "Analysis of the but Diarization System for Voxconverse Challenge," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 5819-5823, doi: 10.1109/ICASSP39728.2021.9414315
21. Caratozzolo, P., Alvarez-Delgado, A., Hosseini, S. (2022). Natural Language Processing for Video Essays and Podcasts in Engineering. In: Hosseini, S., Peluffo, D.H., Nganji, J., Arrona-Palacios, A. (eds) Technology-Enabled Innovations in Education. Transactions on Computer Systems and Networks. Springer, Singapore. https://doi.org/10.1007/978-981-19-3383-7_1.
22. Soenksen LR, Ma Y, Zeng C, Boussioux L, Villalobos Carballo K, Na L, Wiberg HM, Li ML, Fuentes I, Bertsimas D. Integrated multimodal artificial intelligence framework for healthcare applications. NPJ Digit Med. 2022 Sep 20;5(1):149. doi: 10.1038/s41746-022-00689-4. PMID: 36127417; PMCID: PMC9489871.
23. Y. Zhao, X. Xia and R. Togneri, "Applications of Deep Learning to Audio Generation," in IEEE Circuits and Systems Magazine, vol. 19, no. 4, pp. 19-38, Fourthquarter 2019, doi: 10.1109/MCAS.2019.2945210.
24. M. R. Mahmud, A. Cordova and J. Quarles, "Multimodal Feedback Methods for Advancing the Accessibility of Immersive Virtual Reality for People With Balance Impairments Due to Multiple Sclerosis," in IEEE Transactions on Visualization and Computer Graphics, vol. 30, no. 11, pp. 7193-7202, Nov. 2024, doi: 10.1109/TVCG.2024.3456189.
25. Edison Research: Sound-Data-The-State-of-Audio-in-50-Charts. https://www.edisonresearch.com/wp-content/uploads/2024/06/Sound-Data-The-State-of-Audio-in-50-Charts.pdf