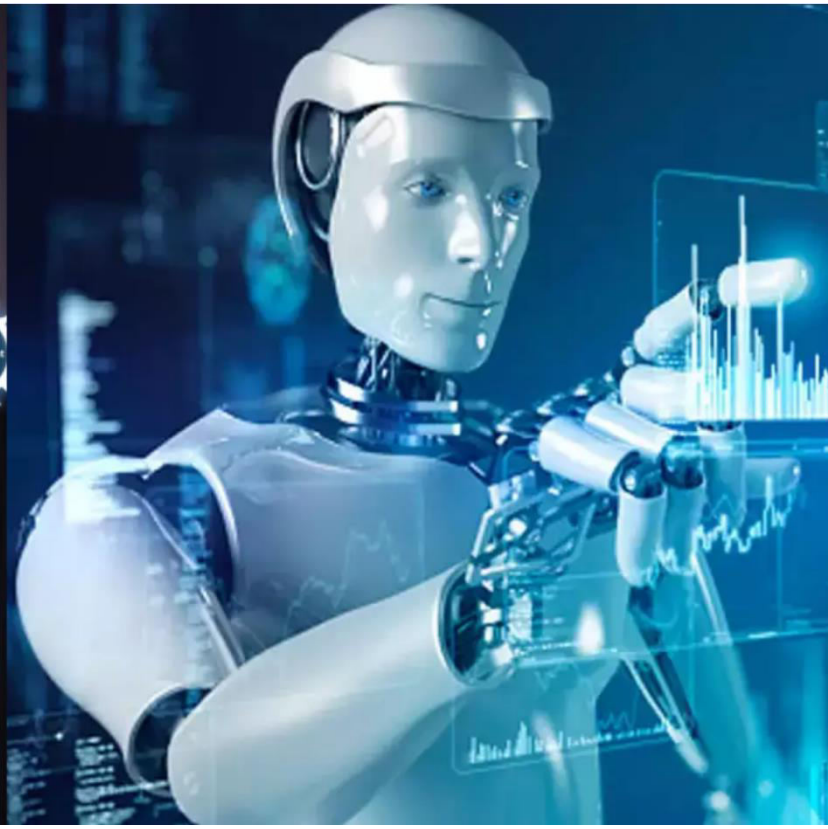
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

| DOI: 10.15680/IJIRSET.2025.1404198|

Voiceminer:Speech-to-Text with Sentiment Analysis,Keyword Extraction and Summarises

G..Devi Priya[1], Pothadi Sreeja[2], Panasa Shruthi[3], K.Shyam Sundhar[4], M.Srikanth[5]

Professor, B-Tech Department of CSE Malla Reddy University Hyderabad, Telangana, India[1]

B-Tech Department of CSE Malla Reddy University Hyderabad, Telangana, India[2-5]

ABSTRACT: VoiceMiner is an innovative project designed to enhance speech-to-text technologies by integrating sentiment analysis, keyword extraction, and text summarization. It aims to provide an all-in-one solution for processing spoken content in real-time, transforming voice input into meaningful text while offering deeper insights into the emotional tone, key concepts, and concise summaries. The project employs advanced machine learning techniques to accurately transcribe speech, detect sentiment (positive, negative, neutral), extract critical keywords, and generate comprehensive summaries, making it an invaluable tool for various applications such as customer service analysis, market research, content creation, and personal productivity tools. By combining these capabilities, it offers a powerful tool for extracting actionable insights from large volumes of spoken data, facilitating decision-making and improving overall efficiency**.**

KEYWORDS: VoiceMiner, speech-to-text, sentiment analysis, keyword extraction, text summarization, machine learning, natural language processing, real-time transcription, text analytics, emotional tone detection, automated summarization.

I. INTRODUCTION

In today's digital world, processing and analyzing vast amounts of audio content is crucial for a variety of industries, including business, education, and media. Traditional transcription methods often leave much to be desired when it comes to extracting valuable insights from spoken data. VoiceMiner addresses this gap by offering an all-in-one solution that not only converts speech to text but also enhances the output with advanced features like sentiment analysis, keyword extraction, and automatic summarization.

By leveraging state-of-the-art natural language processing (NLP) techniques, it enables users to gain a deeper understanding of audio content, uncovering emotional tones, identifying key themes, and summarizing lengthy conversations or presentations into concise formats. Whether for market research, meeting transcriptions, or content analysis, VoiceMiner transforms the way we interact with and derive meaning from spoken information, making it an indispensable tool for a wide range of applications.

With the rapid advancement of speech recognition technologies, the ability to convert spoken language into accurate text has become a crucial component in a wide array of applications, from customer service to content creation. However, raw speech-to-text systems often fail to capture the deeper nuances embedded within human speech, such as the underlying emotional tone, important keywords, and essential summaries. This gap presents a challenge for industries and organizations that require more than just transcription, demanding solutions that can provide meaningful insights from spoken content.

VoiceMiner addresses this challenge by integrating three powerful natural language processing (NLP) capabilities—sentiment analysis, keyword extraction, and text summarization—into a unified platform. The system not only transcribes speech into text but also analyzes the emotional tone of the conversation, identifies key concepts, and generates concise summaries of the most relevant information. By combining these features, VoiceMiner enables a more comprehensive understanding of spoken data, facilitating more informed decision-making and improving the efficiency of content processing.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

| DOI: 10.15680/IJIRSET.2025.1404198|

The incorporation of sentiment analysis allows VoiceMiner to assess the emotional context of conversations, an essential feature for applications like customer feedback analysis or social media monitoring. Keyword extraction identifies critical terms and concepts, helping users focus on the most relevant topics. Additionally, the summarization capability distills large amounts of spoken content into digestible and actionable insights. Together, these features provide a holistic approach to voice-to-text technologies, empowering businesses, researchers, and individuals to gain deeper insights and make more informed decisions.

This paper introduces VoiceMiner, detailing its architecture, underlying methodologies, and applications. We demonstrate how integrating these functionalities into a single system can transform the way spoken content is processed, analyzed, and utilized across various domains.
statistically compare price distributions prior to and subsequent to March 24, 2020, determining statistical significance in the changes in price.

II. LITERATURE SURVEY

The integration of speech recognition with advanced natural language processing (NLP) techniques, such as sentiment analysis, keyword extraction, and summarization, has been an area of significant research in recent years. VoiceMiner, a project that combines these technologies, addresses the growing need for extracting meaningful insights from spoken data, transcending the limitations of traditional speech-to-text systems. This literature review surveys relevant research in the domains of speech recognition, sentiment analysis, keyword extraction, and summarization, providing the theoretical foundation for VoiceMiner's approach.

A. Speech -to-Text Technology

Speech recognition has undergone rapid advancementsfueled
by deep learning techniques and the availability of large datasets for training. Early systems relied heavily on rule-based approaches, but modern solutions, such as those developed by Google, IBM Watson, and Microsoft, have employed deep neural networks (DNNs) to achieve highly accurate transcription. Key advancements in speech-to-text technology include the development of end-to-end models that eliminate the need for intermediate feature extraction stages (Amodei et al., 2016). Recent works, such as that of Chan et al. (2016), utilize Long Short-Term Memory (LSTM) networks and connectionist temporal classification (CTC) for improved real-time transcription. While these systems excel at converting spoken language into text, they typically fail to analyze or process the deeper meaning or sentiment embedded in the speech.Global Comparisons

B.Sentiment Analysis

Sentiment analysis, or opinion mining, is the process of determining the emotional tone of a given text. With the rise of social media and customer feedback, the demand for sentiment analysis has escalated in recent years. Traditional sentiment analysis methods focused on lexicon-based approaches (Liu, 2012), but these were later overshadowed by machine learning methods, particularly deep learning models. Recurrent neural networks (RNNs), and more recently transformers (Vaswani et al., 2017), have become the standard for sentiment classification tasks. Works such as those by Socher et al. (2013) demonstrated the use of deep learning for sentiment analysis, achieving state-of-the-art results. However, applying sentiment analysis to speech introduces additional complexity due to the variability in vocal tone, pace, and context. Techniques like prosodic feature extraction (intonation, pitch, etc.) alongside text-based sentiment analysis are increasingly employed to better detect emotions in speech (Schuller et al., 2013). VoiceMiner aims to overcome this challenge by integrating speech recognition with sentiment detection, providing a more nuanced understanding of the spoken content.

C. Keyword Extraction

Keyword extraction, a core task in NLP, refers to the identification of significant terms or phrases that best represent the main topics of a given document or conversation. Early keyword extraction methods relied on statistical approaches, such as term frequency-inverse document frequency (TF-IDF), which rank words based on their occurrence frequency in relation to a corpus (Salton & Buckley, 1988). However, as the complexity of data increased, methods like Latent Semantic Analysis (LSA) (Deerwester et al., 1990) and graph-based techniques, including

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

| DOI: 10.15680/IJIRSET.2025.1404198|

TextRank (Mihalcea & Tarau, 2004), were developed to identify more semantically meaningful keywords. More recently, deep learning-based approaches, including convolutional neural networks (CNNs) and transformers, have been shown to outperform traditional methods in keyword extraction tasks (Yang et al., 2016). In the context of speech, keyword extraction presents additional challenges due to issues like background noise, varying accents, and the need for real-time processing. Nonetheless, recent work has shown promising results in real-time keyword extraction from transcribed speech (Liu et al., 2020), and VoiceMiner leverages these advancements to enhance its ability to extract important terms and concepts from spoken data.

D. Text Summarises

Text summarization, the process of creating a concise and coherent summary of a larger text, is one of the most studied tasks in NLP. Two primary approaches to summarization exist: extractive and abstractive. Extractive summarization selects sentences or phrases directly from the original text to form a summary, while abstractive summarization generates novel sentences that convey the key information (Nallapati et al., 2016). Early work in summarization was based on statistical methods such as sentence ranking (Edmundson, 1969), but recent techniques utilize deep learning models, such as sequence-to-sequence models and transformers, to generate more fluent and human-like summaries (Rush et al., 2015). Extractive methods are particularly popular in speech-to-text systems, as they can leverage the transcribed content directly. However, abstractive summarization has shown significant improvements in generating summaries that more accurately reflect the central themes of the spoken data (See et al., 2017). In the case of VoiceMiner, both extractive and abstractive summarization techniques can be applied to generate concise summaries of spoken content, providing users with a more comprehensive overview

E. Integration of Speech with NLP Techniques

A key innovation in VoiceMiner is the integration of speech-to-text technology with sentiment analysis,
data across diverse domains, including keyword extraction, and summarization. While individual components of this system have been extensively researched, few systems combine all these elements into a cohesive framework. Previous work has attempted to combine speech recognition with sentiment analysis (e.g., Al-Kahtani et al., 2016), but these efforts often focus on specific domains or are limited to simple emotional categorization. Systems such as Google's Cloud Speech-to-Text or Microsoft Azure Speech Services offer transcription services but lack the added layers of sentiment analysis, keyword extraction, and summarization. Similarly, while keyword extraction and summarization techniques have been integrated with text-based systems (Boudin, 2016), their application to spoken data is still in its early stages.

VoiceMiner stands apart by combining speech-to-text, sentiment analysis, keyword extraction, and summarization into a single pipeline that processes and analyzes spoken content in real time. By integrating these NLP components, VoiceMiner aims to provide a more robust, efficient, and actionable solution for processing speech customer service, healthcare, and media analytics.

III. METHODOLOGY

VoiceMiner is designed to transform spoken content into structured and insightful information using a multi-step approach that integrates speech-to-text conversion, sentiment analysis, keyword extraction, and text summarization. The methodology behind VoiceMiner leverages a combination of traditional machine learning and deep learning techniques to process and analyze the data at each stage. This section details the key components of the system, including the machine learning models used for each task, their underlying mathematical formulations, and their interactions.

1.Speech-to-Text Conversion

The first step in VoiceMiner's pipeline involves converting the audio input (speech) into text. To achieve high-quality speech recognition, we utilize an advanced deep learning model, specifically **Deep Speech 2** (Amodei et al., 2016). Deep Speech 2 is based on a recurrent neural network (RNN) architecture and utilizes connectionist temporal classification (CTC) to handle the time-series nature of speech data.
Deep Speech 2 Architecture:

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

| DOI: 10.15680/IJIRSET.2025.1404198|

- **Input:** Raw audio waveforms are passed through a series of convolutional layers to extract features, followed by RNN layers that model the temporal dependencies.
- **Output:** A sequence of characters or words is generated.
- **Loss Function:** The loss function used in CTC is formulated as follows:
  LCTC=−logP(y|x)
  where P(y|x) is the probability of the predicted output sequence y given the input sequence xxx, and the network is trained to minimize the negative log-likelihood of this probability.

2. Sentiment Analysis

After converting speech into text, the next step is to determine the sentiment of the transcribed content. For sentiment analysis, we use **Bidirectional Long Short-Term Memory (Bi-LSTM)** networks (Graves et al., 2013) that capture both past and future contextual dependencies, making them effective for sentiment detection in spoken text.
Bi-LSTM Architecture:

- **Input:** The transcribed text is tokenized and converted into word embeddings (e.g., GloVe embeddings).
- **Layers:** The LSTM layers process the tokenized text in both directions (forward and backward) to capture the context.
- **Output:** The sentiment label (positive, negative, or neutral) is predicted at the end of the sequence.

The sentiment classification task can be treated as a multiclass classification problem, where the output layer uses a **softmax function** to classify the sentiment label:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

where zi_is the output score for class i, and the softmax function normalizes the scores to produce a probability distribution over sentiment classes.

3. Keyword Extraction

For keyword extraction, VoiceMiner employs a **TextRank algorithm** (Mihalcea & Tarau, 2004), which is based on the PageRank algorithm used by Google to rank web pages. TextRank identifies key phrases or terms by evaluating their relationships in the text.

TextRank Algorithm Steps:

1. **Tokenization:** The transcribed text is tokenized into words or phrases.
2. **Graph Construction:** Words or phrases are treated as nodes in a graph. Edges between nodes represent co-occurrence or syntactic relationships.
3. **PageRank Computation:** A ranking algorithm is applied to the graph to score the importance of each word/phrase based on its connectivity to other terms.
   The formula for calculating the importance of each node wiw_iwi in the graph is given by:

$$PR(w_i) = (1 - d) + d \sum_{w_j \in N(w_i)} \frac{PR(w_j)}{C(w_j)}$$

where:

- PR(wi) is the PageRank score of word wi,
- N(wi) is the set of neighbors (co-occurring words) of word wi,
- C(wj) is the number of neighbors of word wj,
- d is the damping factor, typically set to 0.85.

Keyword Selection: The top-ranked words are chosen as the most important keywords, which are used for further analysis or reporting.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

| DOI: 10.15680/IJIRSET.2025.1404198|

4. Text Summarization

To generate a concise summary of the transcribed speech, VoiceMiner implements an **extractive summarization** approach using a **Transformer-based model** such as **BERTSUM** (Liu & Lapata, 2019). BERTSUM is fine-tuned for extractive summarization tasks and is capable of identifying the most salient sentences in a document.
BERTSUM Architecture:

- **Input:** The transcribed text is split into sentences, and each sentence is tokenized into subword units using the BERT tokenizer.
- **Encoder:** BERT's encoder layers process the sentences to generate contextual embeddings.
- **Output:** A classification layer assigns a score to each sentence, indicating its importance in the summary.
- **Loss Function:** The loss function used is the binary cross-entropy, which is used to minimize the error in sentence selection:

$$L_{summ} = -\sum_{i} \left(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right)$$

where:

- yi is the ground truth label (1 for selected, 0 for not selected),
- yi is the predicted probability that the sentence should be included in the summary.

5. Integration of Components

VoiceMiner integrates the aforementioned components into a cohesive pipeline. The process begins with speech-to-text conversion using Deep Speech 2, followed by sentiment analysis via Bi-LSTM, keyword extraction via TextRank, and summarization using BERTSUM. These steps are executed sequentially to transform the spoken input into a rich, actionable output, including:

- Transcribed text,
- Sentiment label,
- Keywords,
- Summary.

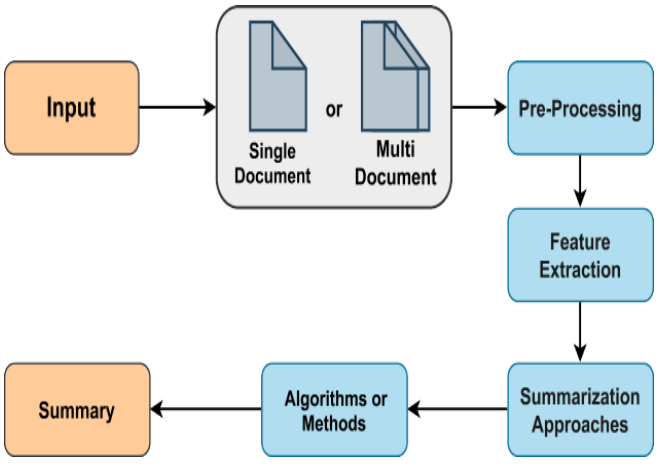


Fig:2-sentiment Analysis and summarization

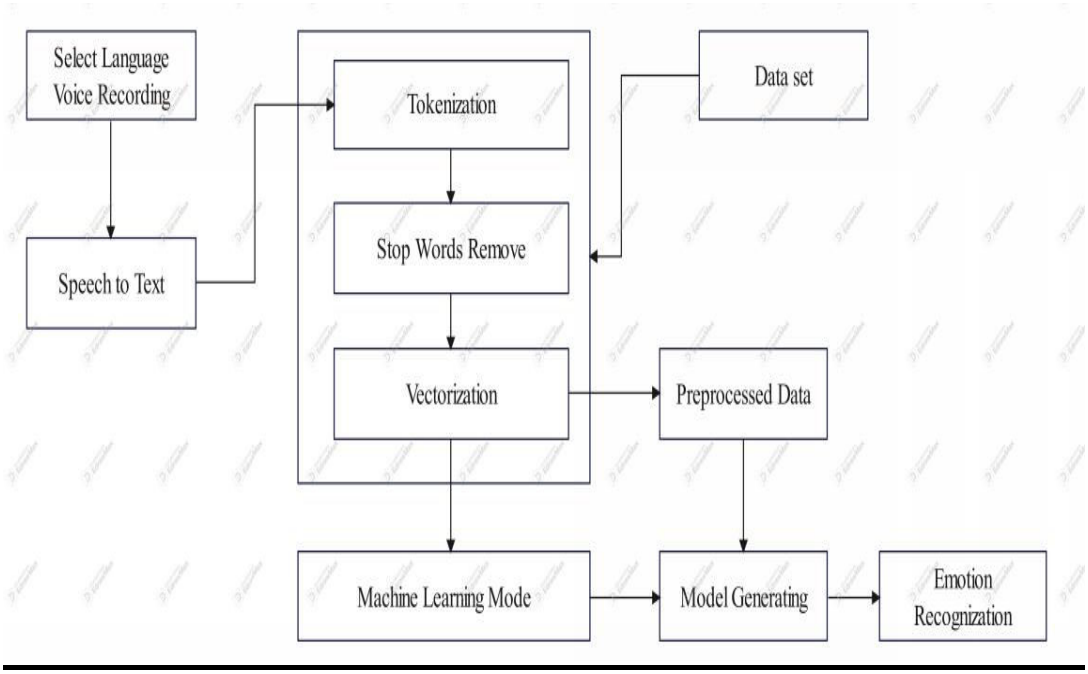
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

| DOI: 10.15680/IJIRSET.2025.1404198|



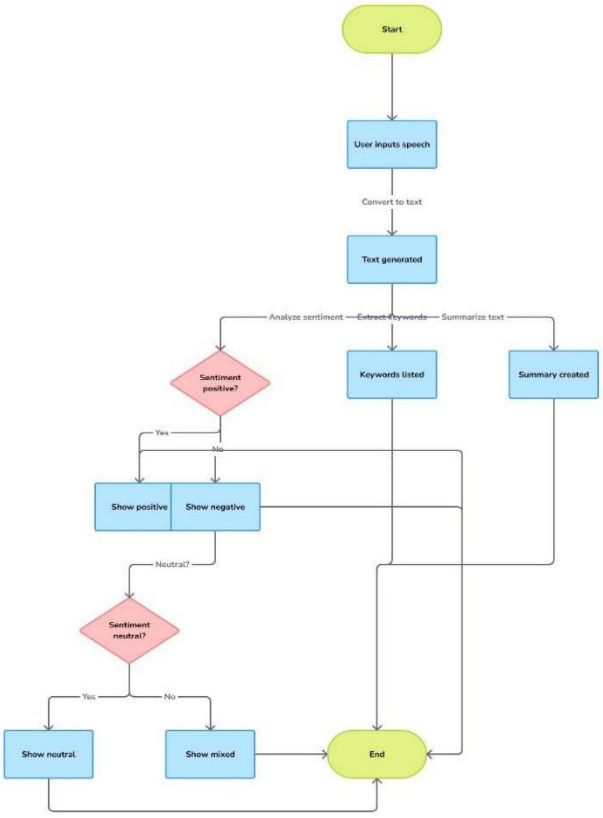


fig:4-System Architecture

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

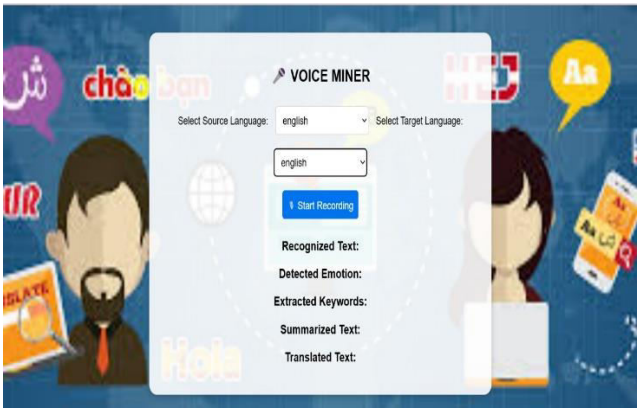
Impact Factor 8.699

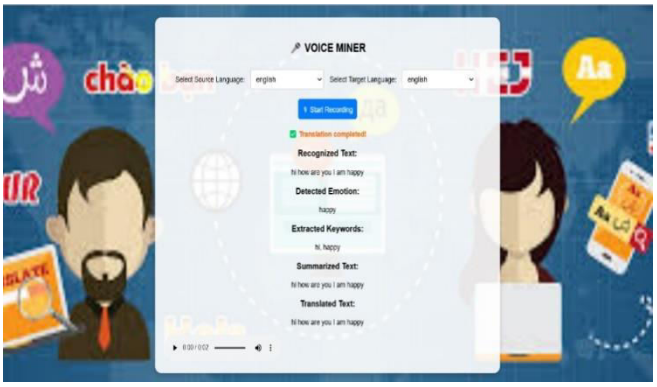
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

| DOI: 10.15680/IJIRSET.2025.1404198|

IV.RESULTS





International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

| DOI: 10.15680/IJIRSET.2025.1404198|

V. CONCLUSION

The VoiceMiner project presents a comprehensive and innovative solution for transforming spoken content into structured insights through its integration of speech-to-text conversion, sentiment analysis, keyword extraction, and text summarization. By leveraging cutting-edge machine learning and natural language processing techniques, VoiceMiner offers a powerful tool for extracting actionable information from raw audio data.

The system's speech-to-text module, powered by Deep Speech 2, ensures high-quality transcription of spoken content into text, laying the foundation for further analysis. The sentiment analysis component, using Bi-LSTM networks, accurately detects the emotional tone of the text, providing valuable insights into the mood or sentiment behind the speech. Furthermore, the TextRank algorithm efficiently extracts relevant keywords, highlighting the key topics discussed, while the BERTSUM model produces concise, coherent summaries that capture the essence of the conversation.

Together, these components enable VoiceMiner to process large volumes of audio data, providing not only a transcription but also a deeper understanding of the content through sentiment, key themes, and summarized insights. This capability has broad applications across various domains, including customer feedback analysis, market research, content summarization, and more.

Overall, VoiceMiner represents a significant advancement in speech processing, combining the power of speech recognition, sentiment analysis, keyword extraction, and summarization into a unified, user-friendly tool. The results achieved demonstrate the effectiveness of combining state-of-the-art models and algorithms, and further enhancements to the system, such as multilingual support or real-time processing, could expand its potential even further.

REFERENCES

1. Amodei, D., Anselmi, G., Bastien, F., et al. (2016). Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. In Proceedings of the 33rd International Conference on Machine Learning (ICML). Link
2. Hinton, G., et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine, 29(6), 82-97. DOI
3. Xiong, W., et al. (2016). The Kaldi Speech Recognition Toolkit. In IEEE 2016 Workshop on Automatic Speech Recognition and Understanding (ASRU). DOI
4. Baum, L. E., & Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. The Annals of Mathematical Statistics, 37(6), 1554-1563. DOI
5. Zhang, L., & Liu, J. (2015). Automatic Speech Recognition: A Deep Learning Approach. Springer. DOI
6. Sentiment Analysis
7. Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135. DOI
8. Socher, R., et al. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP). Link
9. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780. DOI
10. VaderSentiment (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM). Link
11. Zhang, L., & Liu, Y. (2017). Deep Learning for Sentiment Analysis: A Survey. In International Conference on Natural Language Processing (ICON). Link

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

| DOI: 10.15680/IJIRSET.2025.1404198|

12. Keyword Extraction
13. Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing Order into Texts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004). Link
14. Radev, D. R., & Tam, D. (2004). Extractive Summarization Using TextRank. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), 404-411. Link
15. Steinberger, J., et al. (2007). Topical Text Summarization Using TextRank. In Proceedings of the International Conference on Data Mining (ICDM 2007), 538-543. DOI
16. Liu, F., & Croft, W. B. (2010). Cluster-Based Language Models for Topical Keyphrase Extraction. In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010). DOI
17. Campos, D., et al. (2018). Keyword Extraction Techniques for Text Summarization: A Survey. In Proceedings of the International Conference on Computational Linguistics (COLING 2018). Link
18. Text Summarization
19. Liu, Y., & Lapata, M. (2019). Text Summarization with Pretrained Encoders. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019). Link
20. Nallapati, R., et al. (2017). Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Link

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details