IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204224 |

Speech Denoising and Transcription to Improve Audio Communication using Deep Learning

Dr. Nirmala C R [1], Prof. Vishwanath V K [2], Akshay M Kalghatgi [3], Harshavardhan S Shet [3], Kunal Saxena [3], Mohammed Jahangeer [3]

Head of Department, Department of CS&E, Bapuji Institute of Engineering and Technology, Davanagere, Karnataka, India[1]

Assistant Professor, Department of CS&E, Bapuji Institute of Engineering and Technology, Davanagere, Karnataka, India[2]

U.G. Student, Department of CS&E, Bapuji Institute of Engineering and Technology, Davanagere, Karnataka, India[3]

ABSTRACT: In the ever-evolving landscape of call centres, the need for effective and efficient communication is paramount. To meet this challenge, we present a pioneering solution that leverages state-of-the-art audio processing and artificial intelligence technologies. Our system is designed to enhance audio recordings, distinguishing speech from noise components to ensure crystal-clear conversations. Additionally, it excels in classifying speakers within the audio, identifying Speaker 1 and Speaker 2, while seamlessly transcribing the content into text. This multifaceted approach transforms call centre operations, offering a holistic toolkit for optimizing customer interactions. By enhancing audio quality, simplifying the identification of speakers, and providing transcriptions, our system empowers call centre professionals to make informed decisions, monitor compliance, and analyze interactions with precision. The result is a tangible enhancement of productivity, customer satisfaction, and overall operational excellence, positioning this system as a game-changer in the call centre industry.

KEYWORDS: Deep Learning, Short-Time Fourier Transforms, Convolutional Neural Network (CNN), Speaker Diarization, Automatic Speech Recognition (ASR).

I. INTRODUCTION

In an interconnected world, the demand for reliable audio processing and speech recognition systems has never been higher. These systems are vital for extracting valuable information from audio data, even in challenging conditions with background noise and multiple speakers. Our proposed system is a sophisticated and multi-faceted solution designed to address the complexities of audio analysis and transcription. It excels in enhancing audio quality, distinguishing speech from noise, identifying individual speakers, and converting audio content into accurate text transcriptions. With a primary focus on the customer service sector and call centers, our system recognizes the pivotal role of clear communication in the success of businesses. It significantly improves the quality of recorded customer interactions, ensuring that essential conversations are heard with unparalleled clarity. The system's ability to classify speakers within the audio is invaluable when multiple parties are engaged in conversation. It provides clear identification of individual speakers, enhancing transcription accuracy and simplifying the analysis of call center interactions. This feature aids supervisors and managers in monitoring performance and compliance effectively. Our system empowers call center operators to better understand, manage, and optimize their operations, leading to improved customer interactions, increased efficiency, and ultimately, greater success in today's highly competitive business environment.

II. LITERATURE SURVEY

[1] Gökay Dişken et al. (2023) discusses differential convolutional networks for noise mask estimation. It then describes how differential convolutional networks can be used to create noise masks. The article also details the experimental setup and results. The results show that the proposed method outperforms other methods on both clean and noisy data. Key points to take away are:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204224 |

• Noise masks are a type of image processing technique that can be used to remove noise from images.
• Differential convolutional networks are a type of neural network that can be used to learn the relationship between different parts of an image.

• The proposed method uses differential convolutional networks to create noise masks that are more effective than other methods at removing noise from images.

• The proposed method was evaluated on both clean and noisy data, and it outperformed other methods on both datasets.

[2] Daniel Michelsanti et al. (2021) provides a comprehensive overview of deep learning-based approaches for audio-visual speech enhancement and separation.

The authors begin by introducing the challenges of speech enhancement and separation, particularly in noisy environments. They then discuss the advantages of using audio-visual information for these tasks, as visual cues can provide complementary information to improve speech processing.

The paper then delves into the various deep learning architectures that have been employed for audio-visual speech enhancement and separation. These include convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their variants, such as long short-term memory (LSTM) networks. The authors discuss the strengths and limitations of each architecture and how they have been adapted for specific tasks.

The paper also highlights the importance of data preparation and feature extraction for deep learning-based approaches. The authors discuss various techniques for extracting relevant features from audio and visual signals, such as Mel-frequency cepstral coefficients (MFCCs) and spectrograms.

[3] Sebastian Braun et al. (2021) explores efficient speech enhancement techniques using neural networks, particularly RNNs and CRNs. The goal is to identify efficient network architectures that achieve comparable performance to larger models while reducing computational complexity.

The study investigates the influence of RNN size, type, and the use of disconnected parallel RNNs. For CRNs, it examines the convolution layers, spectral vs. Spectro-temporal convolutions, and skip connections.
To ensure generalizability, training is conducted on large-scale data simulating real-world conditions, including reverberation, multiple speakers, various noise types, and varying microphone signal levels. A data generation and augmentation pipeline is proposed for handling reverberant and non-reverberant speech.

[4] Asri Rizki Yuliani et al. (2021) presents an overview of recent deep learning-based methods, including DNN, DAE, RNN-LSTM, CNN, and GAN, aimed at addressing speech enhancement. It evaluates these methods, highlighting the potential for further research in this field. The key observation is that there has been a notable shift in the approach to speech enhancement from using cepstral or spectral representations (time-frequency domain) to waveform representations (time domain). CNNs are preferred due to their ability to reduce computational complexity through convolution and pooling operations and parameter sharing.

[5] Arian Azarang et al. (2020) discusses Speech denoising as a crucial technique for enhancing speech quality and intelligibility in noisy environments. Conventional methods like Wiener filtering and spectral subtraction have limitations in nonstationary noise conditions. Deep learning-based methods have emerged as promising alternatives, demonstrating superior performance in such scenarios.

[6] Jiaming Cheng et al. (2020) proposes a novel DNN-based model for speech enhancement, leveraging self-attention mechanisms on the feature dimension to make optimal use of frame-level information. This model employs two enhancement strategies: firstly, it combines various features like MFCC, AMS, RASTA-PLP, cochleagram, and PNCC to form a 136-dimensional feature representation, enhancing information extraction from different domains. Secondly, a feature-level self-attention mechanism is applied to the output of the fully connected layers, facilitating the capture of internal feature correlations and reducing redundancy. This model outperforms other neural network-based algorithms while using fewer context frames, showcasing its effectiveness in noise suppression and generalization for varying noise scenarios.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204224 |

[7] Virginia Bazán-Gil et al. (2019) discusses the potential of speech technologies, such as automatic speech recognition (ASR) and speaker diarization, for enhancing the accessibility and searchability of television archives. The author highlights the challenges and opportunities associated with applying these technologies to large-scale audiovisual collections. Key points to take away are:

• ASR can be used to generate transcripts of audiovisual content, making it searchable by keywords and phrases.

• Speaker diarization can identify and segment different speakers in audiovisual recordings, facilitating speaker-based retrieval and analysis.

• Challenges in applying speech technologies to television archives include dealing with noisy audio, overlapping speech, and diverse accents and speaking styles.

[8] Aonan Zhang et al. (2019) introduce a novel speaker diarization system that differs from traditional clustering approaches by utilizing a trainable unbounded interleaved-state RNN (UIS-RNN) as the core component. This system is designed for scenarios with access to high-quality training data containing time-stamped speaker labels. On the NIST SRE 2000 CALLHOME benchmark, the new online algorithm outperforms the state-of-the-art spectral offline clustering algorithm when using the same speaker embeddings. An intriguing avenue for future work is to use acoustic features directly, instead of pre-trained embeddings, as the observation sequence for UIS-RNN, thereby creating an end-to-end speaker diarization model.

[9] Dong Wang et al. (2019) presents a comparative analysis of various Automatic Speech Recognition (ASR) methods, specifically CTC-based models, RNN-transducer, and attention models. These models were evaluated in experiments using 80-dimensional log-Mel features as input and a shared vocabulary of characters. The experiments used five bi-directional long short-term memory (BLSTM) layers for encoding and a beam search decoding method. Results indicated that end-to-end models had varying levels of performance, with attention-based models outperforming CTC-based models. However, they still faced challenges compared to hybrid models (HMM-DNN) because of differences in language knowledge learning capabilities.

III. PROPOSED SYSTEM

- Input Data: The system takes call centre conversations, which include both customer and agent speech as well as noise, as its input. These audio recordings are typically noisy and require enhancement.
- Pre-processing: The audio data undergoes pre-processing. This step involves Voice Activity Detection (VAD) to remove gaps and preparatory steps like resampling and trimming for consistency.
- Feature Extraction: Relevant features are extracted from the audio data to represent its spectral and temporal characteristics. STFT (Short-Time Fourier Transform) features are extracted from pre-processed audio, capturing its time-frequency representation essential for sound recognition tasks.
- Denoising: A model trained on paired noisy and clean speech data is utilized to effectively denoise the audio, leveraging statistical properties of features like STFT, spectral, and temporal characteristics.
- Diarization and Transcription: Advanced algorithms for speech recognition and speaker diarization (Unsupervised clustering) are applied. First the speakers in the audio are segregated, and then the speech is converted to text for each speaker respectively, thus enhancing transcription accuracy.

The main objectives of our proposal are:

- To Achieve Speech and Noise Discrimination
- To Implement Speaker Identification
- To Improve Transcription Accuracy

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204224 |

- To Adapt to Call Centre Environment
- To Improve Operational Efficiency

IV. METHODOLOGY

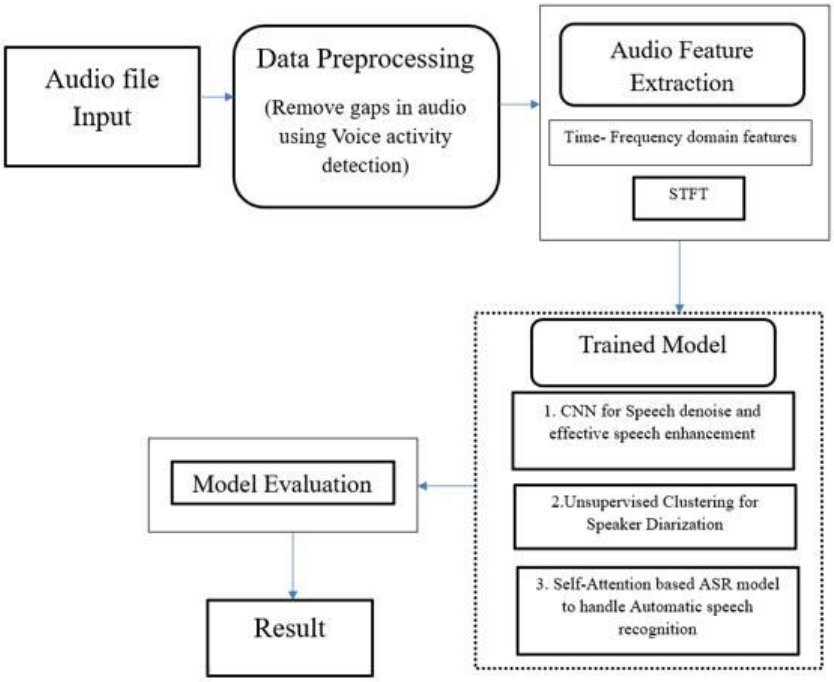

Fig (a) Methodology Diagram

An audio file is taken at the start of the procedure. Voice Activity Detection (VAD) is used in data preparation to eliminate audio gaps from the audio file. Other forms of data pre-processing are also done like resampling audio files to a consistent sample rate and trimming of the audio to a consistent duration to prepare the audio for further analysis. All this is necessary to make the data accurate, consistent, and suitable for analysis. Following pre-processing, the pre-processed audio is utilized to extract STFT features, which capture the time-frequency representation of the audio signal. This representation mimics the human ear's non-linear perception of sound and approximates the response of the human auditory system. Consequently, STFT features are widely utilized in sound recognition tasks. Subsequently, a model trained on a substantial dataset of paired noisy and clean speech recordings is employed to effectively remove noise from the input audio. The model leverages the statistical properties of features, which may include STFT features, spectral features, and temporal features, to distinguish between speech and noise. Advanced speech recognition and speaker diarization algorithms are then employed to transcribe the audio into text. The flow then leads to Model Evaluation, where we can learn how successful the model was for accurately converting the given audio file to text. The performance of the system is evaluated using objective metrics such as signal-to-noise ratio (SNR), perceptual evaluation of speech quality (PESQ) and word error rate (WER). The success of the speech recognition technique depends on the quality and diversity of the training data, the chosen feature extraction methods, the complexity and accuracy of the deep learning model, and the pre-processing and post-processing techniques employed.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204224 |

V. RESULTS

The following chapter presents the results derived at the end of the flow.

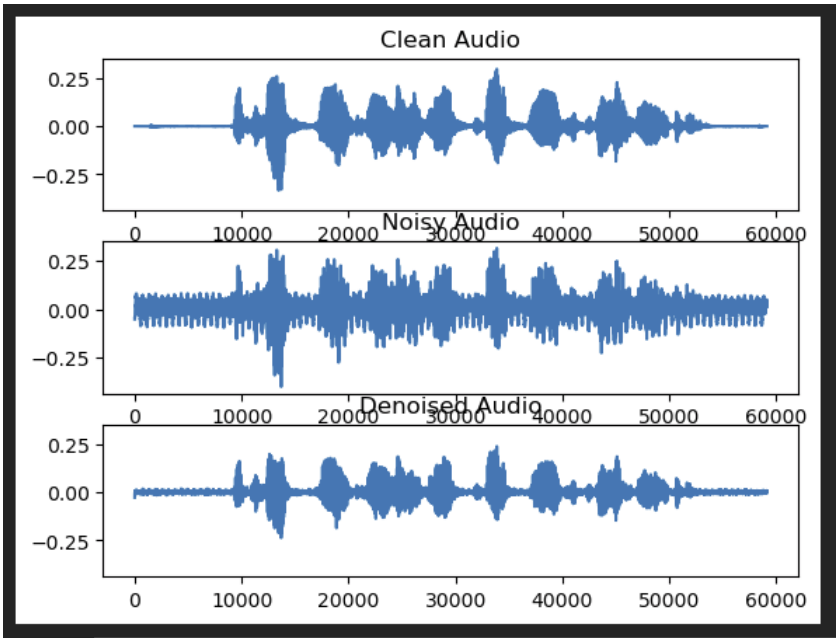

Fig (b) Visualization of audio files

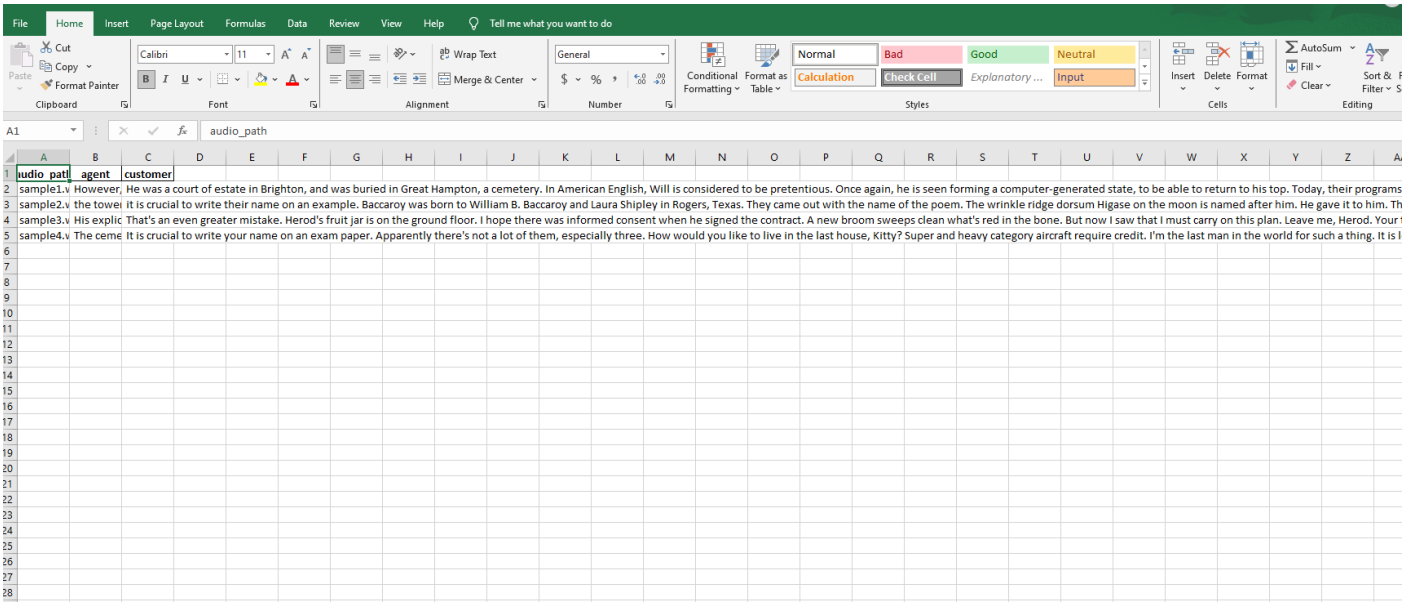

Fig (c) Final Output in Excel

VI. CONCLUSION AND FUTURE SCOPE

In conclusion, the proposed software solution is a ground-breaking advancement in the call centre technology landscape. By incorporating cutting-edge deep learning techniques for noise suppression, Automatic Speech Recognition (ASR), and speaker diarization, this software offers a holistic approach designed to redefine the customer

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204224 |

experience. A standout feature of this software is its remarkable ability to significantly reduce background noise. This solution adeptly identifies and mitigates unwanted noise in recorded audio, ensuring that customer-agent interactions are consistently clear. The software's remarkable speaker diarization capability empowers call centres to accurately identify and monitor individual speakers, a valuable asset for quality assurance and compliance monitoring. The primary output of this software is high-precision, machine-generated text. This text-based output streamlines data storage and analysis, providing call centres with a cost-effective and efficient means to enhance the quality of the services, optimize operations, and elevate customer satisfaction.

In envisioning the future scope of our project, several key avenues for enhancement emerge, poised to elevate the utility and inclusivity of our solutions across diverse applications. Firstly, we aim to pioneer real-time speech denoising and transcription algorithms, enabling seamless integration into platforms like video conferencing, virtual assistants, and live transcription services. This advancement promises to revolutionize communication by ensuring clarity and accuracy, even amidst challenging audio environments. Furthermore, our commitment to inclusivity drives the extension of our models to support multiple languages, fostering global accessibility and engagement. Additionally, our focus on adaptation to various environments empowers our technology to thrive in dynamic settings, from bustling urban landscapes to remote outdoor expanses and even underwater scenarios. Moreover, we are dedicated to enhancing accessibility features, catering to users with speech impediments or accents, thereby democratizing access to our tools and enriching the user experience for all. Through these endeavors, we embark on a journey to redefine the boundaries of communication technology, embracing innovation, inclusivity, and adaptability as cornerstones of our vision for the future.

REFERENCES

1. Dişken, G. (2023). Differential convolutional network for noise mask estimation. Applied Acoustics, 211, 109568.
2. https://www.sciencedirect.com/science/article/abs/pii/S0003682X23003663
3. Michelsanti, D., Tan, Z. H., Zhang, S. X., Xu, Y., Yu, M., Yu, D., & Jensen, J. (2021). An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1368-1396. https://www.researchgate.net/publication/343825817_An_Overview_of_Deep-Learning-Based_Audio-Visual_Speech_Enhancement_and_Separation
4. Yuliani, A. R., Amri, M. F., Suryawati, E., Ramdan, A., & Pardede, H. F. (2021). Speech enhancement using deep learning methods: A review. Jurnal Elektronika dan Telekomunikasi, 21(1), 19-26.
5. https://www.researchgate.net/publication/354252395_Speech_Enhancement_Using_Deep_Learning_Methods_A_Review
6. Braun, S., Gamper, H., Reddy, C. K., & Tashev, I. (2021, June). Towards efficient models for real-time deep noise suppression. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 656-660). IEEE. https://arxiv.org/abs/2101.09249
7. Azarang, A., & Kehtarnavaz, N. (2020). A review of multi-objective deep learning speech denoising methods. Speech Communication, 122, 1-10.
8. https://www.sciencedirect.com/science/article/abs/pii/S0167639319304686
9. Bazán-Gil, V., Lleida, E., Pérez, C., Gómez, M., & Prada, A. (2019). Tecnologías del habla: nuevas oportunidades para los archivos de televisión.
10. http://eprints.rclis.org/38447/
11. Cheng, J., Liang, R., & Zhao, L. (2020). DNN-based speech enhancement with self-attention on feature dimension. Multimedia Tools and Applications, 79, 32449-32470. https://link.springer.com/article/10.1007/s11042-020-09345-z
12. Wang, D., Wang, X., & Lv, S. (2019). An overview of end-to-end automatic speech recognition. Symmetry, 11(8), 1018. https://www.mdpi.com/2073-8994/11/8/1018
13. Zhang A., Wang, Q., Zhu, Z., Paisley, J., & Wang, C. (2019, May). Fully supervised speaker diarization. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp.6301-6305).IEEE.https://arxiv.org/abs/1810.047

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details