Optimizing Audio Clarity: Speech Segmentation and Noise Reduction via VAD and Spectral Subtraction

Suryamritha M, Varshini Balaji, Peeta Basa Pati
Department of Computer Science and Engineering,
Amrita School of Computing, Bengaluru, Amrita Vishwa Vidyapeetham, India
suryamritha.manoj@gmail.com bl.en.u4aie21139@bl.students.amrita.edu, bp_peeta@blr.amrita.edu

***Abstract*—In many audio processing applications, the occurrence of noise reduces the overall quality of the signal. This affects the performance of the subsequent steps such as Speech recognition, and speaker identification. This paper presents a real-time noise reduction and segmentation approach for audio signals using spectral subtraction. The proposed method segments the audio. Segment-wise noise reduction is beneficial in terms of computational efficiency and flexibility. The algorithm used is Voice Activity Detection (VAD) the algorithm can achieve real-time performance and handle large audio files efficiently by processing smaller segments individually. It also allows easy parameter tuning and optimization based on each segment's characteristics, hence enhancing noise reduction's effectiveness. It is effective as it also removes pauses, silences and unnecessary sounds making the audio contain only relevant information. The efficiency of the proposed method is evaluated by comparing various metrics such as Root Mean Square (RMS) energy, Amplitude, Zero Crossing Rate (ZCR) spectral contrast, and spectral flatness.**

***Index Terms*—noise reduction, segmentation, spectral subtraction, parameter tuning, Voice Activity Detection, RMS energy, Amplitude, ZCR, spectral contrast, spectral flatness.**

I. Introduction

Voice activity detection (VAD) is a critical preprocessing step for speech-related activities like automated speech recognition (ASR) [1]. Recent research indicates that voice enhancement techniques can supplement VAD systems [2]. Speech enhancement tries to enhance the overall perceived quality of speech signals utilising various audio signal processing techniques [3]. The loud environment reduces the performance of the voice recognition system [4]. Speech enhancement is difficult due to the variety of background noise sources. The majority of existing approaches focus on simulating speech rather than noise [5].

Segmentation plays a crucial role in noise reduction. Segmentation isolates the segments using various techniques based on n certain criteria, such as energy level, zero-crossing rate, or other statistical properties. Segmentation helps in the removal of silent and paused regions in the audio. It ensures that only relevant audio segments are combined in the final output.

This adds to improved performance and quality in noise reduction applications. It plays an essential role in speech recognition systems. Speech segmentation is important in various fields such as Natural Language Processing (NLP) and Speech recognition.

It helps us understand language semantics, identifying individual words or phrases and improving language language comprehension. It helps in machine translation by identifying the boundaries of phrases or sentences. Also using speech segmentation, we not only segment data but it plays a crucial role in removing silences, pauses, and other non-meaningful data. It only combines meaningful data for subsequent processing. This process increases the efficiency and accuracy of tasks such as transcription, speech recognition, and speaker diarization, as it ensures that the focus remains on the content-rich portions of the audio.

The algorithm used for segmentation is VAD. VAD is a crucial component in speech processing systems as it identifies segments of audio that contains speech. It helps to distinguish between periods of speech and non-speech. It's analyzed in short frames. Overall the study is about noise reduction using spectral subtraction and segmentation using VAD that can be used in further processing of various applications that are related to speech. Subsequent sections follow a structured format: Section II presents a literature review, Section III details methodology and contributions, Section IV covers results and analysis, and Section V concludes, summarizing findings and suggesting future research directions.

The key contributions of this paper are :

- Usage of VAD algorithm to extract meaningful segment containing speech and spectral subtraction for noise reduction in speech
- Contributes by enhancing audio quality by removing silences and keeping only relevant information.

.

II. LITERATURE REVIEW

The practice of breaking up continuous speech into more manageable, meaningful parts, such words or phonemes, is known as speech segmentation. It is required for many speech-processing applications, including natural language understanding and automatic recognition of voices. Segmentation techniques vary from rule-based approaches to more complex machine learning algorithms.

Dinler et al. [6] have developed new methods for Kurdish language processing, using gated recurrent unit (GRU) RNN for Kurdish speech segment recognition. They developed a Kurdish-specific database and optimized processing settings. The research developed an extensive Kurdish vocabulary dataset for consonant, vowel, and silence (C/V/S) discrimination-based segment identification. The study used hybrid feature vector approaches to accurately characterize phoneme boundaries and improved C/V/S discrimination, highlighting the importance of hybrid characteristics, window types, and classification models.

Kürzinger et al. [7] combines publicly accessible German voice corpora, amounting to more than 1700 hours of data, some of which contain unlabeled speech. Using a pre-trained ASR model based on Connectionist Temporal Classification (CTC), more training data is first extracted from non-segmented or non-labeled sources in this two-step data preparation process. Utterances are then extracted using probabilities of labels generated by the CTC trained network to build segment alignments. Using these improved training data, we aim to train a hybrid CTC attention Transformer model.

Guo et al. [8] studied the relationship between speaking style and speech segmentation, finding that clear speech has better acoustic-phonetic clues than conversational speech, improving segmentation. However, hyperarticulated, clear speech may not always result in greater segmentation due to decreased coarticulation. The study also found that in quiet settings, clear speech facilitated segmentation, but not in noisy ones. Baevski et al. [9] developed wav2vec-U, a voice recognition model that uses self-supervised speech representations to segment unlabeled audio. The model outperformed earlier unsupervised techniques, reducing phone mistake rates and achieving a WER of 5.9 on the English Librispeech benchmark.

Bredin et al. [10] proposes an all-encompassing segmentation model that incorporates voice activity detection, overlapped speech recognition, and speaker change detection for speaker diarization. Our model treats this as a classification of multiple label issues with permutation invariant training, drawing inspiration from EEND. It runs at a high temporal resolution (every 16 ms) on 5-second audio chunks. According to experiments conducted on several datasets, VAD and overlapped speech detection have significantly improved. Additionally, it can be used as a post-processing step to assign to speeches that overlap.

Federico et al. [11] have improved a speech-to-speech pipeline, enabling automatic dubbing.Neural text-to-speech adapts the duration of each utterance; prosodic alignment of the translation with the original speech segments; audio rendering enhances text-to-speech output by removing background noise and reverberation from the original audio; and neural machine translation generates output of desired length. We present a subjective assessment of the apparent naturalness of automatic dubbing and the relative merits of every suggestion for enhancement for the English-to-Italian TED Talk segments.

Chen et al. [12] introduce the concept of continuous speech separation (CSS) to produce non-overlapped speech signals from continuous audio streams with partially overlapped utterances. They present a new dataset called LibriCSS, which simulates conversations using far-field microphones and was created from LibriSpeech. The evaluation process evaluates the performance of a speaker-independent CSS algorithm using a Kaldi-based ASR protocol. Kreuk et al. [13] describe a self-supervised representation learning algorithm that works directly with raw waveforms to find phoneme boundaries using an unsupervised method. The model trains independently without phonetic transcriptions or user comments, showing superior performance in evaluations using TIMIT and Buckeye corpora. Experiments on three languages and unseen distributions show the advantages of adding more untranscribed data.

Shajeesh et al. [14] use compressive sensing to minimize impulsive noise in voice signals, comparing it to spectrum subtraction, total variation denoising, and signal-dependent rank order mean algorithm. Gowri et al. [15] use Variational Mode Decomposition and ℓ_1 trend filter to improve voice signals damaged by white Gaussian noise, surpassing techniques like Spectral Subtraction and Minimized MSE.

Kumar et al. [16] propose using DL attention-based approaches to understand and recognize spoken emotions, exploring CNN-LSTM and Mel Spectrogram-Vision Transformer (ViT) models. Their experimental results support the extraction of features technique of DL-based approaches, reducing the need for manual feature selection in ML classifiers. Chowdary et al. [17] aim to create models for dysarthria diagnosis and data comprehensibility using a few-shot technique based on a transformer model to address previous examinations' data leaks.

Poorna et al. [18] developed a weight-based emotion detection system for classifying emotions from audio signals from South India. They constructed an audio directory containing five emotional states and tested it for subjective validity. Key characteristics were retrieved and categorization methods like KNN, SVM, and Neural Networks were used. The system was compared to methods using usual, weighted, and feature combinations. Yan et al. [19] proposed a graph signal processing theory for speech enhancement, introducing the graph spectral subtraction technique to decrease noise interference in noisy speech. The iterative graph spectral subtraction approach was also proposed for voice augmentation performance.

Haider et al. [20] describes a novel method for denoising respiratory sounds utilising empirical mode decomposition (EMD), Hurst analysis, and spectral subtraction. The proposed denoising approach might be a valuable tool for doctors in producing unambiguous analyses of respiratory sounds. Future research will concentrate on separating heart sounds from respiratory sound signals. accurate identification and segmentation.

Hence, VAD is a crucial step in speech segmentation, identifying regions of speech presence within an audio signal. Spectral subtraction, a common method in VAD, works by

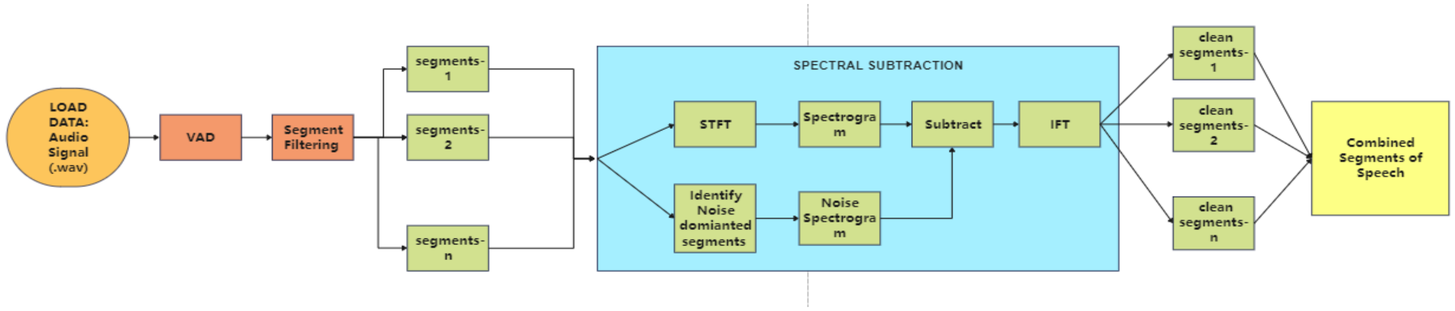

Fig. 1. Proposed Methodology

estimating and subtracting noise from the input signal to enhance the clarity of speech segments, aiding in their

III. METHODOLOGY

The proposed methodology for noise reduction using spectral subtraction by the using segmented speech signals which was obtained using VAD is depicted in Fig 1. The final output has large silent sections removed using silence thresholds and enhanced audio quality.

A. Segmentation

VAD is an essential preprocessing step in speech and audio processing tasks. Its objective is to identify segments of an audio signal containing meaningful speech or crucial information. It identifies non-silent intervals based on a specific threshold. VAD techniques analyze temporal characteristics such as amplitude, frequency, duration of speech segments, pauses between words or sentences, and the overall rhythm to determine if speech is present or not.

B. Segment filtering

Once the segments are identified, we further refine them. The duration of each segment is examined, and short-lived segments are filtered out as they do not contain any meaningful content. So overall from the original audio,long pauses, silence, and other irrelevant data are also removed. This process ensures that only significant portions of the audio are retained, improving the overall quality and coherence of the output.

C. Noise reduction

Spectral Subtraction technique is used for noise reduction. The segmented audio is the input that's taken and Spectral Subtraction is computed to obtain its magnitude spectrum. The source speech is in time domain. The signal is divided into short overlapping segments to analyze the frequency components of the signal over short intervals of time. Then Fourier transform is computed for each segment. As a result, the signal is converted from the time to frequency domain using Short-time Fourier Transform (STFT). The magnitude spectrum obtained from the STFT demonstrates how energy is dispersed across distinct frequency bands over time. The noise spectrum is calculated by taking the average of the magnitude spectra from many frames in the noise-only areas. This provides a representative estimate of the background noise present in the signal. Once the noise spectrum has been predicted, it is subtracted from the magnitude spectrum of the entire signal. Then subtraction operation is performed to attenuate the background noise present in the signal while preserving the spectral components corresponding to the speech or desired signal. The clean segmented audio is reconstructed using inverse STFT.

D. Adjustable Parameters

It is necessary to ensure that spectral subtraction has proper parameter tuning, such as choosing the size of analysis windows, the flooring threshold, and the method for noise estimation, for achieving optimal noise reduction performance. The threshold parameter used in VAD refers to the threshold level in decibels (dB) above which an interval in the audio signal is considered non-silent. This can be adjusted to control the sensitivity of the VAD algorithm. Higher threshold value identifies segments with relatively high energy levels as speech while lower values capture a broader range of audio content as speech.

E. Visualization

To provide insights into the audio data, the waveform and spectrogram of the original and the final audio is plotted. The waveform provides information about the temporal aspects of the signal, whereas the spectrogram gives information about its frequency content.

IV. RESULTS

The implementation uses spectral subtraction to reduce noise in a voice activity detection and segmentation system. The method effectively distinguishes speech from background noise by using spectral subtraction to derive noise profiles and then segmenting the audio depending on observed voice activity. The successful segmentation procedure is demonstrated by playing and visualizing the divided audio snippets. Fig. 2 shows the Original Audio waveform and its corresponding spectrogram. Fig. 3 shows the Combined Speech Segments Audio waveform which is the final output after combining the necessary segments and its spectrogram. It can be inferred that in Fig. 3 all the silence and noisy parts are removed compared

TABLE I
COMPARISON OF THE INITIAL AND FINAL AUDIO BASED ON SILENCE

Audio	Silent Frames	Silence Duration (sec)	Silence (%)	Total Duration (sec)
Original	96521	2.01	15.33	13
Final	2140	0.04	1.26	3

TABLE II
COMPARISON OF THE ORIGINAL AND FINAL AUDIO

Metric	Original Audio	Final Audio
RMS Energy	0.00518	0.0149
Amplitude (dB)	-41.15	-22.98
ZCR	0.0898	0.0357
Spectral Contrast	16.34	20.91
Spectral Flatness	0.0143	0.00016

to Fig. 2 and the duration of the final combined segments audio is reduced.

Table I demonstrates how efficiently our proposed methodology works by removing noise and silences by comparing silent frames, duration, and the silence percentage. The final audio has significantly reduced the unnecessary silences. Features like Root Mean Square (RMS) energy, amplitude in decibels (dB), and Zero Crossing Rate (ZCR), are calculated for analyzing the silent frames. Then, it defines thresholds for these features to identify frames with low audio activity (potential silence) based on RMS, dB, and ZCR. Finally, it combines these frames and calculates the total silence duration (frames divided by sampling rate) and silence percentage relative to the entire audio file.

Table II shows the comparison of both the original and final audio based on various metrics.The RMS energy of final audio is higher, signifying that the audio is now louder and has more energy or power. This rise of loudness can be beneficial and help make the audio louder and more high energetic and full sounding overall. Also, the final audio also has a low negative amplitude dB value, which definitively indicates that the present audio has a high level of loudness.Zero Crossing Rate measures the rate of sign changes in the audio signal, which correlates with high-frequency noise. The original audio has a higher ZCR with a value compared to the final audio. This shows that the original audio has more high-frequency noise components compared to the final audio. A lower ZCR indicates fewer high-frequency noise components. The spectral contrast in the final audio points to greater variability between spectral peaks and valleys, implying that it contributes to a perception of richer and more detailed sound. .Furthermore, the final audio example has a significantly lower value of spectral flatness, which indicates that the audio is less noise-like and has more tonal components. It is often associated with a louder and more pleasant timbre as compared to other equally distributed harmonic sounds. Hence The final audio exhibits higher RMS energy, less negative dB values, higher spectral contrast, and lower spectral flatness, indicating a louder, more dynamic, tonally richer, and less noisy sound compared to the original audio.

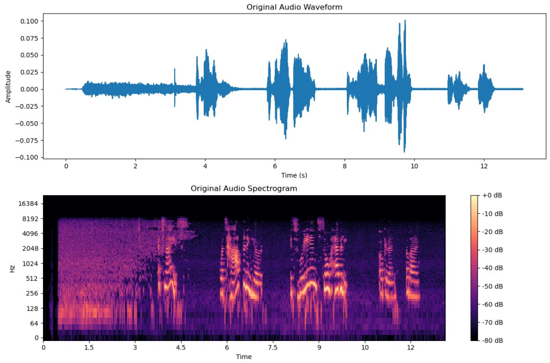

Fig. 2. Original Audio waveform and it's spectrogram

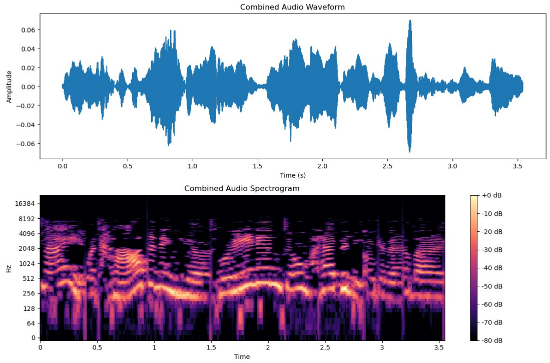

Fig. 3. Combined Speech Segments Audio waveform and it's spectrogram

VAD is essential in speech segmentation as it helps identify segments of audio containing speech, enabling accurate boundary detection. Spectral subtraction complements VAD by effectively reducing background noise in speech signals, enhancing the clarity of speech segments, and improving the accuracy of segmentation algorithms by distinguishing speech from noise. Combined, VAD and spectral subtraction contribute to more precise and efficient speech segmentation processes.

V. CONCLUSION AND FUTURE WORK

The VAD and segmentation system that has been put into place shows how effective spectral subtraction is in separating speech segments from noisy surroundings. Through precise detection of voice activity and use of noise reduction algorithms, the system improves the comprehensibility and sharpness of the separated audio. These findings highlight the possibility of using techniques based on spectrum subtraction to enhance speech processing in noisy environments. Segmenting speech signals enhances noise removal performance by enabling analysis and adjustment of thresholds for individual segments, thereby optimizing noise reduction techniques. This approach offers flexibility and adaptability to address varying noise levels within specific segments, leading to improved overall speech quality and segmentation accuracy.

To improve noise reduction performance in a variety of audio situations, future improvements may concentrate on fine-tuning the spectral subtraction parameters. Further investigating and integrating various noise reduction methods into the segmentation pipeline may also enhance the resilience and dependability of the system. Moreover, the system's capacity to adjust to changing noise profiles and speech features may be improved by using machine learning models for contextual information-based adaptive noise reduction and segmentation.

References

[1] H. Dinkel, S. Wang, X. Xu, M. Wu and K. Yu, 2021. ”Voice activity detection in the wild: A data-driven approach using teacher-student training,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, pp.1542-1555.

[2] X. Tan and X. L. Zhang, 2021, June. ”Speech enhancement aided end-to-end multi-task learning for voice activity detection,” In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6823-6827). IEEE.

[3] V. R. Balaji, S. Maheswaran, M. B. Rajesh, M. Kowsigan, E. Prabhu, and K. Venkatachalam, ”Combining statistical models using modified spectral subtraction method for embedded system,” Microprocessors and Microsystems, vol. 73, p. 102957, 2020.

[4] A. S. Dhanjal and W. Singh 2024. ”A comprehensive survey on automatic speech recognition using neural networks,” Multimedia Tools and Applications, 83(8), pp.23367-23412.

[5] C. Zheng, X. Peng, Y. Zhang, S. Srinivasan and Y. Lu, 2021, May. ”Interactive speech and noise modeling for speech enhancement,” In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 16, pp. 14549-14557).

[6] Ö. B. Dinler and A. Nizamettin. ”An optimal feature parameter set based on gated recurrent unit recurrent neural networks for speech segment detection,” Applied Sciences 10, no. 4 (2020): 1273.

[7] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll. ”Ctc-segmentation of large corpora for german end-to-end speech recognition,” In International Conference on Speech and Computer, pp. 267-278. Cham: Springer International Publishing, 2020.

[8] Z. C. Guo, and R. Smiljanic. ”Speaking clearly improves speech segmentation by statistical learning under optimal listening conditions,” Laboratory Phonology 12, no. 1 (2021).

[9] A. Baevski, W. N. Hsu, A. Conneau, and M. Auli. ”Unsupervised speech recognition,” Advances in Neural Information Processing Systems 34 (2021): 27826-27839.

[10] H. Bredin, and A. Laurent. ”End-to-end speaker segmentation for overlap-aware resegmentation,” arXiv preprint arXiv:2104.04045 (2021).

[11] M. Federico, R. Enyedi, R. Barra-Chicote, R. Giri, U. Isik, A. Krishnaswamy, and S. Hassan. ”From speech-to-speech translation to automatic dubbing,” arXiv preprint arXiv:2001.06785 (2020).

[12] Z. Chen, Y. Takuya, L. Liang, Z. Tianyan, M. Zhong, L. Yi, W. Jian, X. Xiong, and L. Jinyu. ”Continuous speech separation: Dataset and analysis,” In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7284-7288. IEEE, 2020.

[13] F. Kreuk, K. Joseph, and A. Yossi. ”Self-supervised contrastive learning for unsupervised phoneme segmentation,” arXiv preprint arXiv:2007.13465 (2020).

[14] K. U. Shajeesh and Dr. K. P. Soman, ”Noise Cancellation Method for Robust Speech Recognition,” International Journal of Computer Applications (IJCA), 2012.

[15] G. B. Gowri and Dr.K. P. Soman, ”Enhancement of white Gaussian noise affected speech using VMD-ℓ_1 trend filter method,” in *Journal of Intelligent and Fuzzy Systems*, 2018, vol. 34, pp. 1701-1711.

[16] C. S. Kumar, D. M. Ayush, M. K. Advaith, S. S. Srinath, H. Sannidhi, G. L. Jyothish, and R. Vinayakumar. ”Speech Emotion Recognition Using CNN-LSTM and Vision Transformer,” In *International Conference on Innovations in Bio-Inspired Computing and Applications*, pp. 86-97. Cham: Springer Nature Switzerland, 2022.

[17] P. N. Chowdary, M. S. Akshay, V. S. Aravind, M. S. Aashish, G. V. N. S. L. V. Vardhan, and G. Jyothish Lal, ”A Few-Shot Approach to Dysarthric Speech Intelligibility Level Classification Using Transformers,” in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India, 2023, pp. 1-6, doi: 10.1109/ICCCNT56998.2023.10308067.

[18] S. S. Poorna, K. Anuraj, and G. J. Nair, ”A weight based approach for emotion recognition from speech: An analysis using South Indian languages,” In Soft Computing Systems: Second International Conference, ICSCS 2018, Kollam, India, April, 2018, Revised Selected Papers 2 (pp. 14-24). Springer Singapore.

[19] X. Yan, Z. Yang , T. Wang, and H. Guo, 2020. ”An iterative graph spectral subtraction method for speech enhancement,” Speech Communication, 123, pp.35-42.

[20] N. S. Haider, 2021. ”Respiratory sound denoising using empirical mode decomposition, hurst analysis, and spectral subtraction,” Biomedical Signal Processing and Control, 64, p.102313.