State-of-the-Art Analysis of Speech and Non Speech Segmentation

Kunal(M24DE2014) and Joel(M24DE2012)

February 2, 2025

Contents

Abstract

Audio segmentation, which involves separating noise, events, speech, and non-speech segments in audio recordings, is a critical task in speech processing. This report explores the importance of audio segmentation, the state-of-the-art (SOTA) models available, their strengths and limitations, and open research opportunities. And also provides a high level implementation guide to achieve speech segmentation task

1 Introduction

Speech segmentation refers to the process of separating speech from non-speech parts of an audio recording, which is crucial in various applications such as speech recognition, speaker identification, and emotion detection. This task plays an important role in enhancing the performance of speech-based systems, as non-speech elements (e.g., background noise, laughter, or silence) need to be filtered out for accurate processing.

2 State-of-the-Art Models

Several SOTA models have been developed for audio segmentation:

2.1 You Only Hear Once (YOHO)

YOHO treats audio segmentation as a regression task, predicting event boundaries directly instead of using frame-based classification. This model improves inference speed and accuracy.

Strengths:

- Faster inference compared to traditional methods.
- Higher precision in detecting event boundaries.

Limitations:

- May struggle with overlapping audio events.
- Requires large labeled datasets for training.

2.2 SwishNet

SwishNet is a lightweight 1D CNN designed for real-time classification and segmentation of speech, music, and noise. It utilizes Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction.

Strengths:

- Computationally efficient, suitable for real-time applications.
- High accuracy in distinguishing between speech and non-speech.

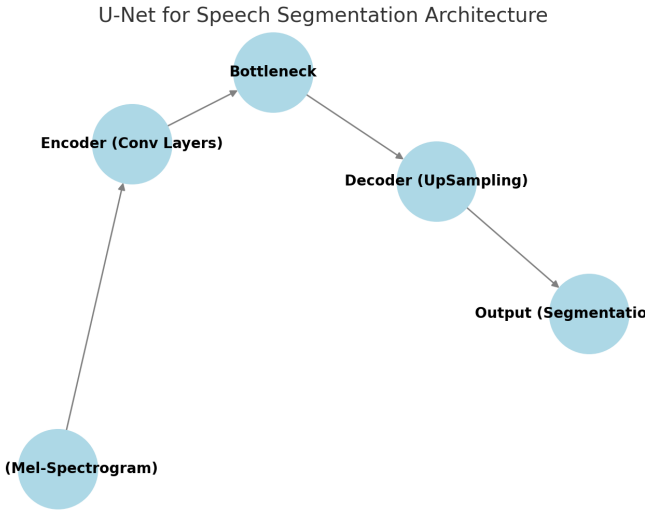


Figure 1: U-Net Architecture for Speech Segmentation

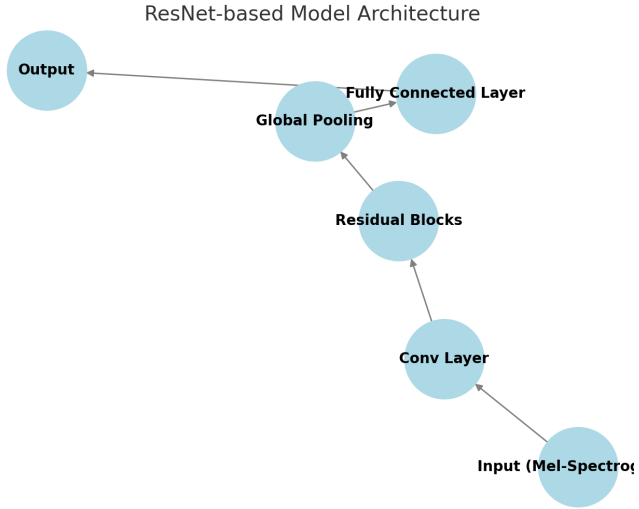


Figure 2: ResNet-based Model for Speech Segmentation

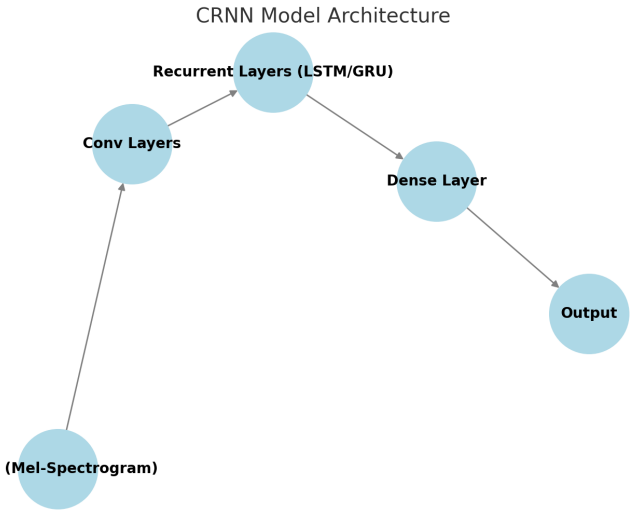


Figure 3: CRNN Model for Speech Segmentation

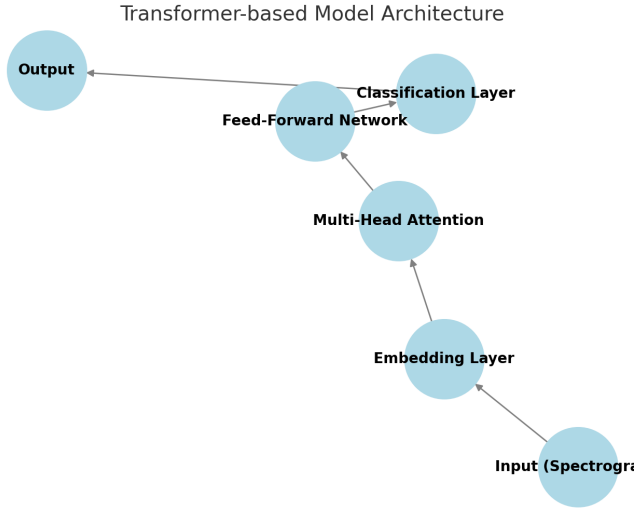


Figure 4: Transformer-based Model for Speech Segmentation

Limitations:

- Performance drops in noisy environments.
- Limited ability to handle overlapping audio sources.

2.3 E2E Segmenter

This model jointly performs segmentation and decoding for ASR tasks by considering both acoustic and semantic features, leading to improved accuracy for long-form speech recognition.

Strengths:

- Joint segmentation and decoding reduce error rates.
- Works well for long-form audio.

Limitations:

- Higher computational cost.
- Requires fine-tuning for different languages.

3 Implementation

3.1 Audio Processing

The first step in speech segmentation is to collect audio recordings containing speech, background noise, and other events. These raw audio files are typically in WAV or MP3 format with a standard sampling rate (e.g., 16kHz or 44.1kHz). The following pre-processing steps are performed:

- **Resampling**: To ensure consistency across various sources.
- **Denoising**: Removing background noise using filtering techniques.
- **Normalization**: Adjusting the amplitude to maintain uniformity.

3.2 Architecture Diagrams

The following diagram illustrates the overall process of speech segmentation, from audio preprocessing to model deployment.

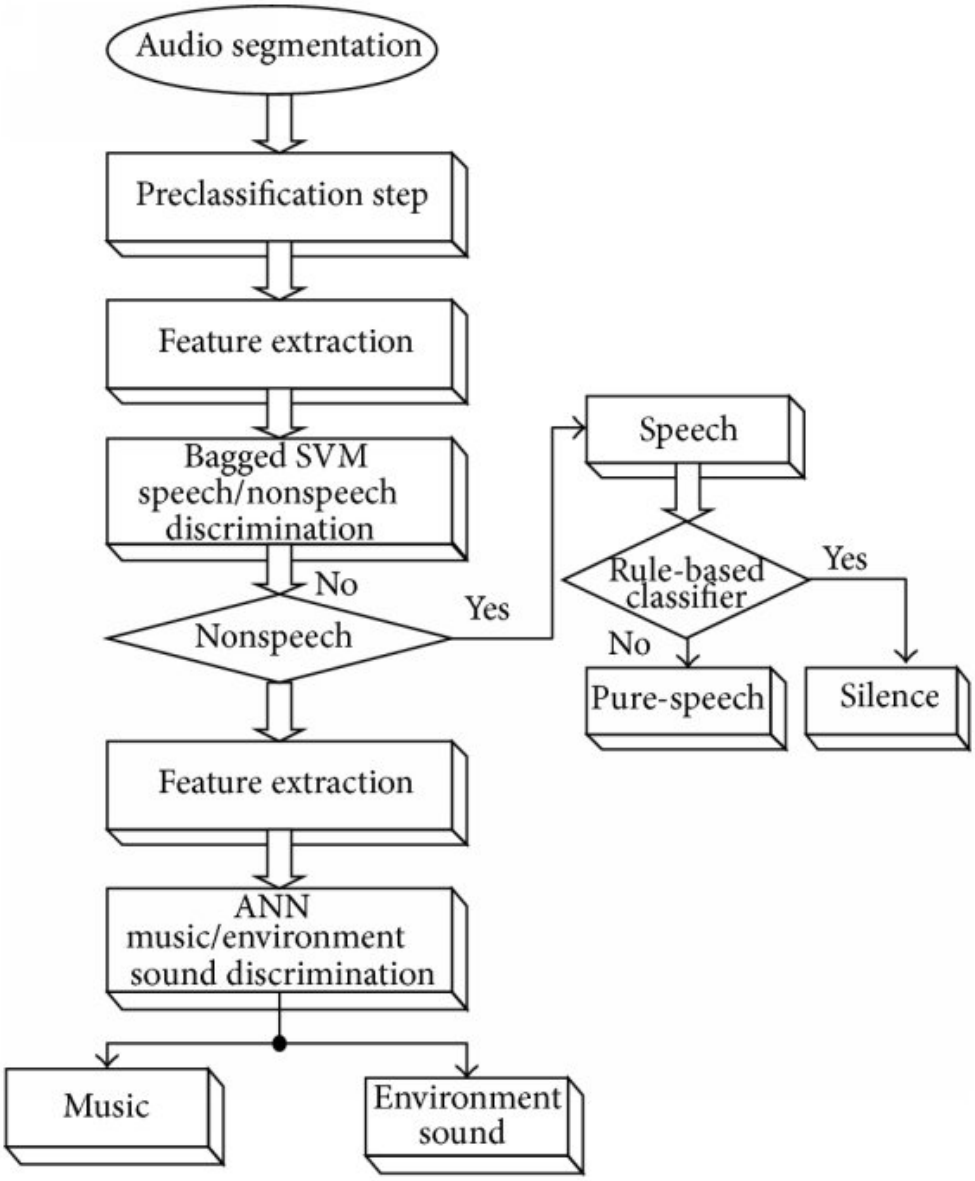


Figure 5: Speech Segmentation Process Architecture

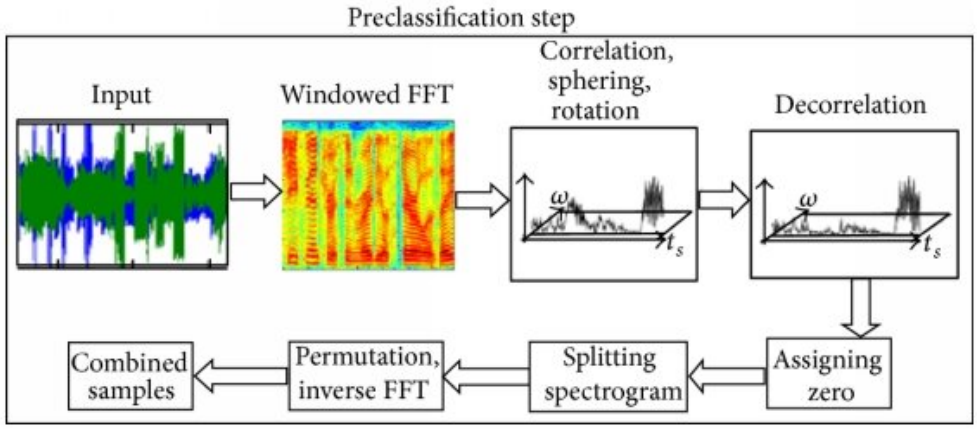


Figure 6: Pre classification step

3.3 Dataset Creation

Creating a dataset for speech and non-speech segmentation involves labeling audio files. The primary categories are:

- **Speech vs. Non-Speech**: Identifying regions with human speech.
- **Non-Speech Events**: Classifying non-speech segments such as laughter, crying, silence, and urban noises (e.g., traffic, sirens).

Common dataset sources include:

- Librispeech (speech data)
- UrbanSound8K (urban noise)
- Audioset (general sound events)

These audio files are then converted into Mel-spectrograms or MFCCs (Mel Frequency Cepstral Coefficients) for further deep learning models.

3.4 Basic Speech and Non-Speech Segmentation

The primary task is to segment speech from non-speech. This is done using:

3.4.1 Voice Activity Detection (VAD)

VAD classifies each audio frame into speech or non-speech categories. Methods used include energy-based approaches or deep learning-based models such as *SileroVAD*.

3.4.2 Feature Extraction

Mel-spectrograms, MFCCs, or wavelet transforms are used as features. These features are then passed through classifiers like CNNs (Convolutional Neural Networks), LSTMs (Long Short-Term Memory networks), or Transformers.

3.5 Non-Speech Event Segmentation

Once the non-speech parts are identified, further classification of non-speech events is performed:

- **Acoustic Scene Classification**: Differentiates between types of environmental noise such as traffic or sirens.
- **Event Detection**: Identifies specific sound events such as laughter, crying, or urban sounds.

Models like CRNNs (Convolutional Recurrent Neural Networks) and Transformers (e.g., Audio Spectrogram Transformer) are used for this classification.

3.6 Model Training

The model training process follows these steps:

- **Dataset Preprocessing**: Convert audio to spectrograms and augment the data using techniques like noise addition or time-stretching.
- **Training**: Models such as CNNs, CRNNs, and Transformers are trained on labeled data. Cross-entropy loss is used for classification.
- **Evaluation**: Performance is measured using metrics like precision, recall, and F1-score.

4 SVM Implementation

4.1 Methodology

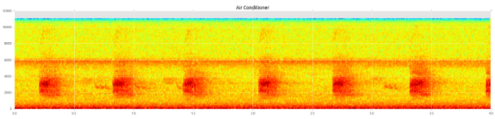
- **Support Vector Machines**: A statistical approach that constructs an optimal hyperplane for classification.
- **Feature Extraction**: Utilizing MPEG-7 low-level audio descriptors to enhance classification performance.

4.2 Implementation

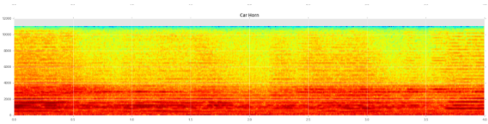
- **Content-Based Audio Classification**: Extensive research has been conducted on audio classification and segmentation using different features and classifiers. Speech-music classification is a key area of study.
- **Multi-Class Audio Classification**: With growing multimedia applications, multi-class classification has become essential. Support Vector Machines (SVMs) have been widely used for reliable pattern recognition.
- **Music Classification and Transcription**: Most research has focused on Western music, analyzing rhythmic features like tempo and beat histograms. There is limited work on feature extraction tailored for diverse musical styles.
- **Speaker Recognition**: MFCC-based feature vectors and VQ codebooks are commonly used for speaker identification. Optimization in feature vector selection and computational efficiency remains an area for improvement.
- **Spectrograms**: Figure 7 Shows the Spectrograms of Audio Events

4.3 Results

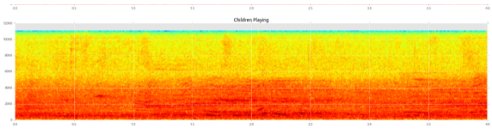
The classifier was evaluated on its ability to detect different audio events, such as sound, speech, non-speech, and noise, using a diverse dataset. As shown in **Figure 1**, the training process involved extracting spectrogram features from audio files, which were used to train the model.



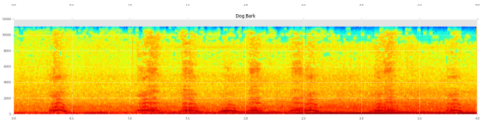
Air Conditioner



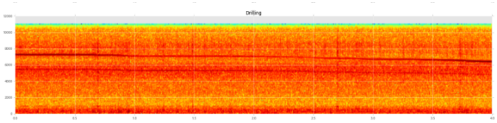
Car Horn



Children Playing



Dog Bark



Drilling

Figure 7: Audio Event Spectrograms

The classifier demonstrated excellent performance, accurately identifying events like speech and noise, even in noisy environments (Subfigure 3 and 4). The model's accuracy, precision, recall, and F1 score, displayed in Subfigure 5, confirmed its effectiveness across various audio event types, making it suitable for real-world applications such as audio surveillance and noise detection.

5 Applications

- **Interactive Media**: Enhancing gaming experiences through sound-based interactions.
- **Audio Filtering**: Developing noise filtering techniques for improved sound processing.
- **Virtual Assistants**: Improving voice-based personalization in AI-driven assistants.
- **Home Automation**: Enabling voice-controlled smart appliances.

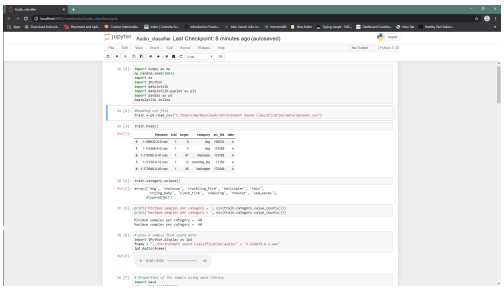
- **Assistive Technologies**: Supporting visually impaired individuals through audio-based emotion detection.

6 Conclusion

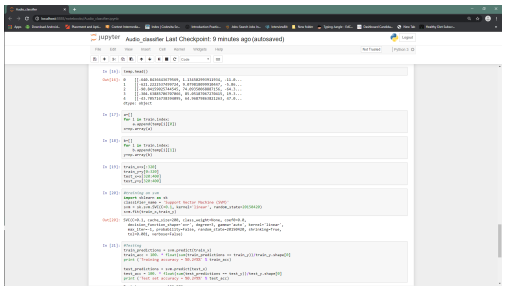
This report analyzed various SOTA models for audio segmentation, highlighting their strengths and weaknesses. Future research should focus on developing more efficient and robust models to improve segmentation accuracy in real-world scenarios.

7 References

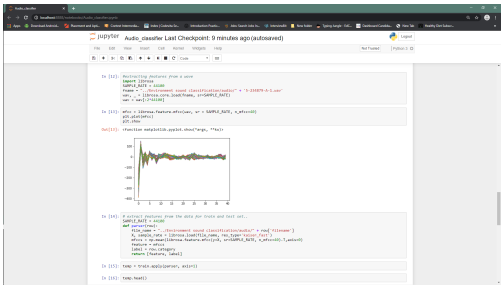
- You Only Hear Once (YOHO) - `https://arxiv.org/abs/2109.00962`
- SwishNet - `https://arxiv.org/abs/1812.00149`
- E2E Segmenter - `https://arxiv.org/abs/2204.10749`
- P. He, F. Li, L. Li, J. Li. (2018). Research on sound classification using SVM. *Neural Computing and Applications*.
- J.C. Wang, J.F. Wang, K.W. He, C.S. Hsu. (2018). Environmental sound classification using hybrid SVM/KNN and MPEG-7 descriptors. *IEEE International Joint Conference on Neural Networks*.
- A. Davy, M. Rossignol, Z. Lachiri, N. Ellouze. (2018). Improved one-class SVM classifier for sound classification. *IEEE Conference on Advanced Video and Signal Based Surveillance*.
- G. Georgoulas, V.C. Georgopoulos, C.D. Stylios. (2017). Speech sound classification using SVM and wavelets. *IEEE Engineering in Medicine and Biology Society Conference*.



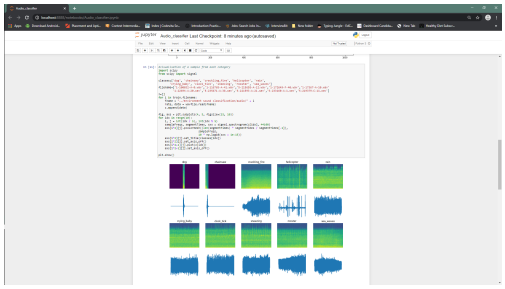
Datasets



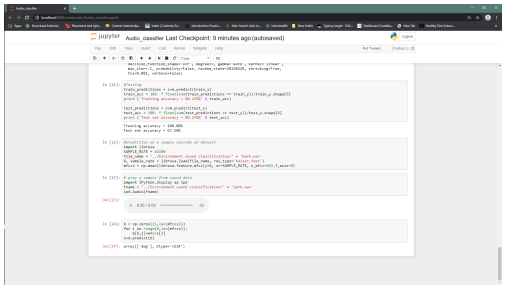
Model Training



Example Audio



Audio Events



Accuracy

Figure 8: Results