Voice Activity Detection for Children's Read Speech Assessment in Noisy Conditions

DDP Stage-II Thesis

submitted in partial fulfillment of the requirements
for the degree of

Master of Technology

by

Ankita Pasad

Roll No: 12D070021

under the guidance of

Prof. Preeti Rao

Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai 400076, India.

June 2017

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Signature)

Ankita Pasad

(Name of the student)

12D070021

(Roll Number)

Abstract

Recordings of read-aloud stories by children in a school setting can be used to provide an assessment of reading skills via automatic speech recognition (ASR) technology. ASR, however, is known to be highly susceptible to background noise. An unusual variety of foreground (breath release, mic pops, etc.) and background (children playing, distinct background talker, wind, etc.) non-speech sounds is found in the recordings obtained via the rural school setting. Motivated by the observation that close to 50% of the recorded audio comprises purely non-speech activity, we investigate robust approaches to voice activity detection (VAD) to eliminate non-speech segments to the extent possible prior to ASR.

We consider features related to the speech signal characteristics which provide a good distinction between speech and noise under different noise settings. A supervised classifier trained on a set of traditionally used features based on energy, pitch, harmonicity, modulation and entropy has been used to obtain the frame-level speech/non-speech decisions. Based on our observations on application-specific data, we synthesize noisy data by adding relevant noise types at different levels to the naturally acquired clean data. The trained model for speech/non-speech classification is tested on both the real-world noisy as well as the synthesized noisy data. The results obtained are evaluated using both frame level and segment level evaluation metrics. The segment level metrics used are known to be correlated with the ASR performance. The ASR results on real-world noisy data processed through VAD have also been reported to show an improvement of about 6% in the phone error rate when compared to the one without any preprocessing. The performance of the proposed VAD is compared against two previous approaches - the VAD proposed for Adaptive Multi-rate, a standard speech codec algorithm and VAD based on the recently proposed single frequency filtering approach.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

It is well known that in India's large rural population, millions of children complete primary school every year without achieving even basic reading standards [4]. More shockingly, the Annual Status of Education Report(ASER) for 2016 [5] reports that the ability to read English has degraded over the years 2007-2016 for upper primary grades and that for lower primary grades it has remained the same. Although the reasons are varied, lack of reliable teaching resources is one of the principal culprits. Automatic literacy assessment, one of the promising applications of speech processing, can prove to be a dependable solution, and can also aid automate and standardize the educational surveys such as ASER which require over 25000 volunteers [5]. If implemented successfully, it will not only make learning more efficient and interactive for the students but will also help lift the burden off teacher's shoulders so that they can concentrate more on planning and can provide individualized help using reliable feedback from the assessment system. The collective goal of the project [6] is to consider a scalable technology solution that facilitates oral reading practice via story-reading tasks and provides a reliable assessment of read speech to the students with limited availability of language teachers. With such a system in place, we aim to not only provide a feedback on the pronunciation accuracy of the reader but also be able to comment on the suprasegmental aspects like comprehension, fluency, confidence. This comes with technical challenges due to well-known difficulties with child speech recognition [7] because of acoustic variabilities, lack of training data and influence of dialect leading to train-test mismatch. Furthermore, high susceptibility of data to various noise sources due to the lack of resources to facilitate a noise-free environment in rural schools adds to the challenges involved in ASR task for this data. In this work, we review the ways to robustly handle noisy data before forwarding it to the ASR engine and propose a voice activity detector (VAD) system which is observed to positively affect the ASR performance.

1.1 Why Voice Activity Detection?

Data collection, discussed in detail in chapter 2, is done in a rural school environment. Such an environment is prone to a variety of noises such as rain, background talker, bell, children playing on the ground, to name a few. Given the lack of human resources to ensure discipline in and around during the recording activity, combined with the unavailability of infrastructure to provide even enough space proportional to the number of students participating, it is nearly impossible to eliminate the presence of noise sources in the acquired recordings. Thus, conducting a streamlined, noise-free recording session is not possible. For instance, at a tribal school

near Mumbai, where tablet based story reading is a scheduled and supervised activity conducted during the school hours for the students in grade 5 to 7, the students record the stories in their recess time. 4-5 students would be seated in a small classroom, as can be seen in figure 1.1, just beside their playground. Simultaneously other kids would be playing just outside the classroom and some curious ones would gather near the window, as in figure 1.2, chit-chatting while looking at the participating kids. At the other site, one of the classrooms is reserved for the reading sessions while the smaller standard kids are in the rooms adjacent to this one, shouting, playing, making all kinds of noise. It is very tough to keep such noise sources under control even with enough teachers to look over them because we simply cannot shun all the activities going on in the background. Apart from this, there are naturally occurring sounds too like rain, wind, bird chirps, which inadvertently interfere with the recording. All-in-all the environment is not very conducive to acquire clean recordings, and processing the collected noisy recordings directly could be very deleterious to the ASR.

Figure 1.1: A reading session

Based on the free-form and un-standardized data collection environment, and the presence of non-speech components there is an obvious need for a speech enhancement system to be in place. Furthermore, it has been empirically observed that these recordings consist of the non-speech segments for almost half of the duration, thus making it more challenging for the ASR as more the presence of non-speech frames, more are the chances of incorrectly flagging it as some phone or the other. Thus, we wish to eliminate the presence of non-speech segments to the extent possible before providing it to the ASR engine. Getting rid of the noise regions and being able to chunk the whole 3 to 6 minutes long audio recordings into smaller speech regions could also aid the recognition accuracy by reducing the possibility of false alarms. Moreover, the separated noise regions could also be used to train the filler model(s) in the ASR system. While transcribing the utterances on word-level, the annotators are also asked to label whether it is clean or noisy. It is found that over 80% of the total duration of the complete annotated dataset ($\sim$31 hrs) is labeled as noisy. At times, the environment is too noisy and/or the reader has a very soft voice, thus rendering the whole recording useless for recognition purposes because of a very low signal-to-noise ratio (SNR). Not being able to process the recordings for providing the feedback

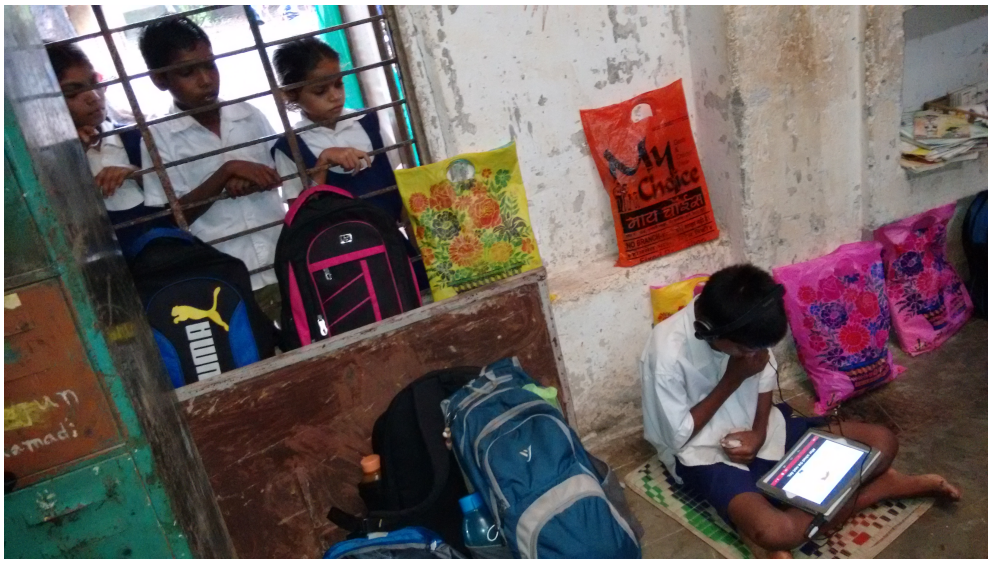

Figure 1.2: Onlookers during one of the reading sessions

after the child has put efforts in reading the story will only demotivate him/her. Thus our aim is to successfully classify the noise regions from noisy speech ones and to be able to flag when the environment is not conducive for recording.

Over the past few years, there has been considerable progress in the field of speech recognition. This has resulted in highly accurate performance for specific tasks in constrained environments. However, their performance degrades rapidly once the environment becomes noisy. The reasons include unaccounted non-speech segments in training which corrupt the acoustic models, non-speech segments in testing which lead to false alarms because of weak filler model. Though there are other measures to deal with noisy data directly at the ASR level [8], these call for a large amount of labeled task-specific data; and given that the ASR system is already fraught with other above-mentioned challenges, we propose to use a robust VAD as a pre-processing step. Although VAD has been extensively studied in the past [9, 10], there is very little work, to the best of our knowledge, which proposes to use VAD as a pre-processing step for an ASR task. For ASR we would also need to take into consideration the types of errors made by VAD, as incorrect clipping of speech by VAD in different regions of a speech segment will have different effects on the ASR accuracy [11]. The proposed system should lead to a significant reduction in the percentage duration of noise segments, while not affecting the speech segments and thus giving an overall improvement in ASR performance.

1.2 Thesis Outline

Dataset, different noise types and related observations are detailed in chapter 2. In chapter 3, various approaches for VAD mentioned in the literature have been discussed, including the aspects of benchmarking the system and the evaluation criteria. In chapter 4 the proposed VAD framework using a classifier-based approach has been discussed, followed by detailed analysis of the results for VAD as a stand-alone system and as an end-to-end system combined with ASR engine in chapter 5. We conclude the thesis by indicating the possible future directions based on our present analysis in chapter 6.

Chapter 2

Database

2.1 Data Collection and Annotation

A reading tutor application has been deployed on a mobile tablet, which provides for a low-cost, portable device that can be easily handled by children. The screen space of a 7-inch tablet is sufficient for the story-reading task with an audio-visual presentation of stories with text. Moreover, children from rural regions are fascinated by technology and show a great enthusiasm in using it for learning purposes as well. We have adopted Sensibol Reading Tutor (RT) app [12], an Android OS application for the project. This app has availability of customization for classroom use with multiple separate child accounts, where the user has to register with his/her photograph and a unique password. All the subsequent activity is then logged into individual user accounts. This setting encourages the user to make mistakes and learn from them, without any hesitation or fear of being watched over or critically judged by the teacher. On logging in, a list of stories is displayed on the screen, as is seen in the figure 2.1. The student can choose from any of the eighteen stories present in the app database. These stories are selected from BookBox [13], a readily available rich resource of illustrated text designed for child readers. On choosing the story, the student can choose from three different modes of usage. In *listen* mode, the audio-visual of the story starts playing along with the story text broken down into appropriate chunks to be displayed at a time. The text gets highlighted as the narrator reads the story aloud, as can be seen in figure 2.2. In *record* mode, the student can read-aloud the selected story with or without any assistance from the narrator audio. The student can listen to the story as many times as required before moving on to this mode. The *review* mode provides an option to revisit the past recordings, where the user can listen to his/her narrations along with the actual ones, synchronized with the video [14]. All recordings are collected at 16kHz sampling rate, using a headset mic to minimize background noise.

As mentioned before the complete story text is broken down into appropriate chunks to be displayed at a time. Similarly, on the annotation interface (designed by SensiBol Audio Technologies [12]) the labeling is to be done sentence-by-sentence, as shown in the figure 2.3. This is for the ease of the transcriber, as then he/she can listen to the corresponding chunks of narrator audio and user-recorded audio, and annotate just one sentence at a time, rather than listening to the whole 3-6 minutes recording at once. Not only this, as mentioned before, the ASR will perform better on a few seconds audio than on the one which is a few minutes long. So, we need to divide the complete story into sentences accurately for the annotation and ASR purpose. This

can be achieved with an aid of the video time-frames information, by making an assumption that the sentence would be read completely while it is visible on the screen. But while the kids are learning to read and speak English, some might not be able to keep up with the pace of the video. There are many cases where kids continue to finish the present sentence while the screen displays the next one. So a system to detect speech activity would come handy here as well. We can then use the video time-frames information with some tolerance in the duration adjustment depending on VAD decision boundaries. The only difference is that for the interface we would like to retain the silences within a sentence as intra-sentence pause durations will affect the suprasegmental (prosodic) ratings, while for ASR tasks we want to remove as many pauses as possible. The overview of the steps involved and where does VAD fit in is shown in figure 2.4. The sentence-level audio chunks are further uploaded on the rating interface and once labeled, they form a part of the training data for ASR.

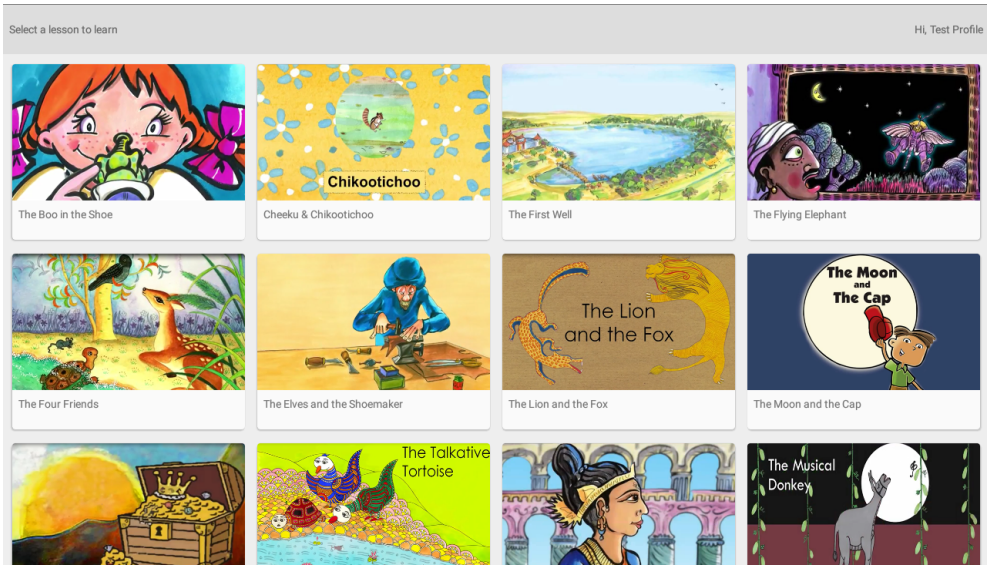

Figure 2.1: Screenshot of SensiBol RT app: The list of stories

2.2 Application-specific Data

This data has been collected as per the procedure mentioned in section 2.1. 28 noisy recordings, each consisting of a complete story, from this database are selected and are manually labeled for speech/non-speech ground truth along with the annotation for noise type. The average duration is 2.75 minutes long and the average speech content is observed to be 47%, which validates that almost half of it is pure noise. These recordings have a fair distribution of various noise types encountered at the recording site. Further details are mentioned in table 2.2. In the classifier-based VAD, this real-world dataset is mainly used for testing purposes.

2.2.1 Observations

Table 2.1 lists different noise types observed along with their respective characteristics. It thus shows that a noise type may vary in multiple aspects like pitch, amplitude, harmonicity, and could be either present for the whole duration or could be intermittent, with presence at isolated locations only. Background talker, where a distinct background speech is clearly heard, is the

Figure 2.2: Screenshot of SensiBol RT app: Highlighted in white is the already spoken text, while the green highlight denotes the upcoming narration.

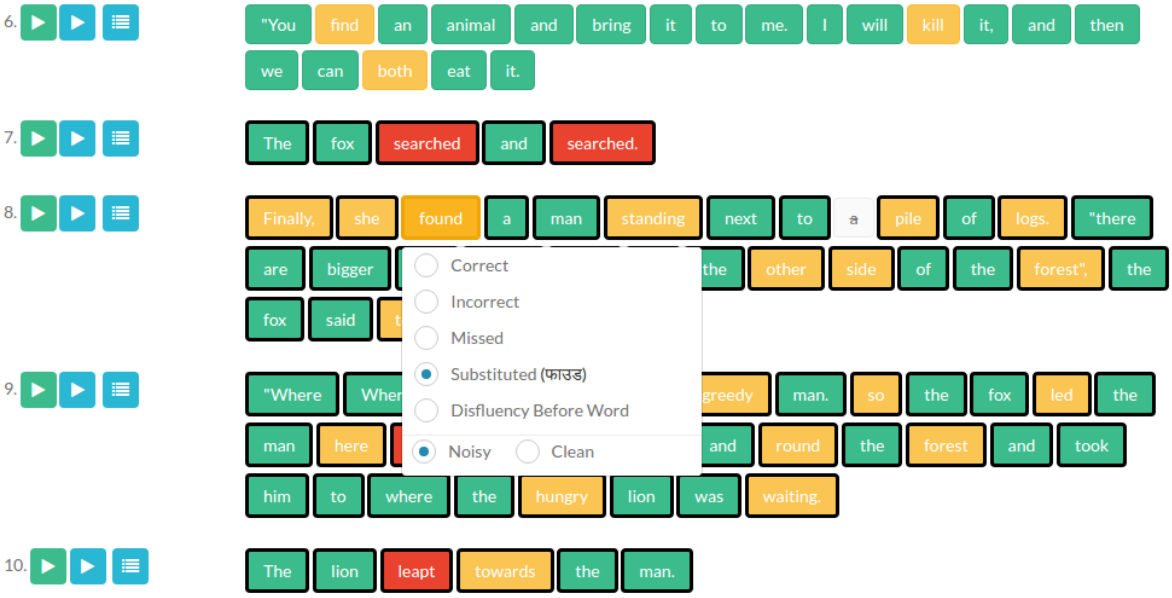

Figure 2.3: Screenshot of the interface designed for annotation, showing a few lines of the transcibed story

most challenging case of all, as it shares all the attributes with speech regions. Breath release, which is quite common, is difficult to deal with as its characteristics resemble that of unvoiced fricative phones. This type of noise can also get accentuated because of the microphone. School noise (further referred to as children playing) is very prevalent as the recordings are made in a school environment; this noise type includes kids shouting, talking, running around, sounds of things being thrown, thumped. Babble, which is another very common sound, constitutes mainly of multiple people indistinctly talking at a distance from the microphone. Rain, generator are the examples of stationary noises, while the latter contains steady harmonics. Mic spurts are quite frequent too. They occur when the kid is so loud that the amplitude gets clipped off by the microphone, or when the speech is very breathy, the breaths get accentuated and degrade

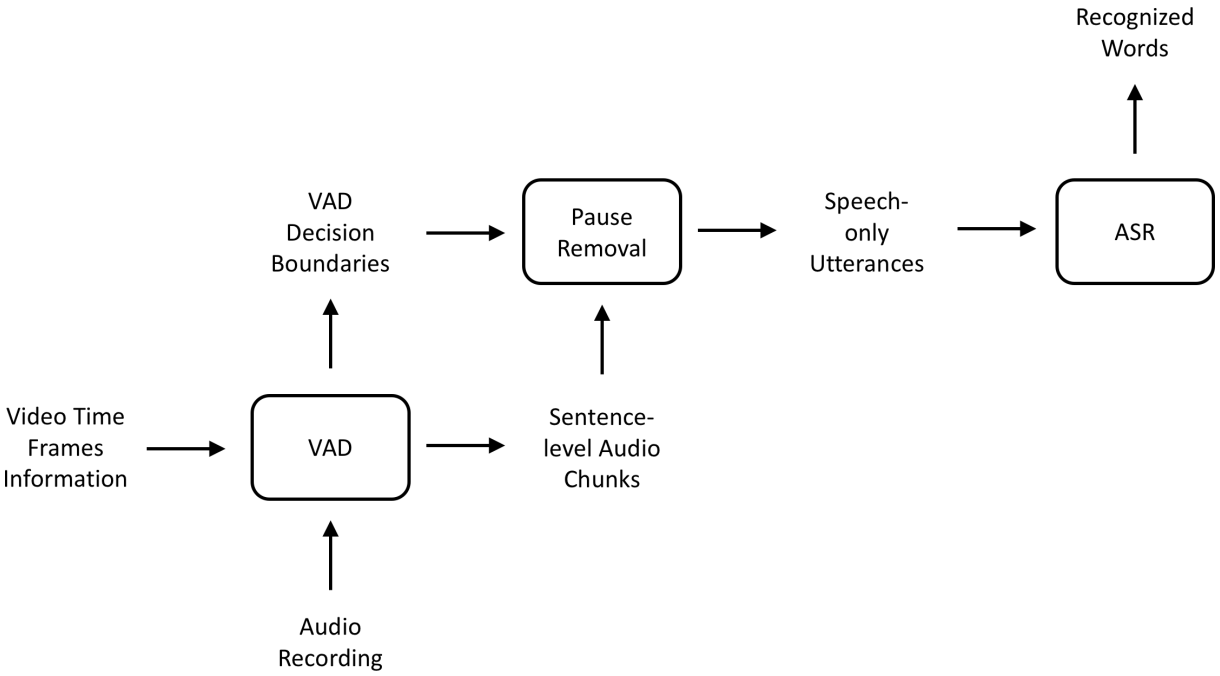

Figure 2.4: The complete overview of the processing steps involved

the speech quality. This type very rarely occurs in non-speech regions, so speech/non-speech classification cannot remove it. Wind is yet another interesting noise type which can greatly degrade the speech regions if a higher amplitude noise region overlaps with speech.

Noise type	**Characteristics**	**Occurrence**
Background talker	Speech-like	Might or might not be pervasive
Bell	High pitch and stationary	Intermittent
Rain	White noise-like spectral characteristics	Pervasive
Mic noise	Dense spectrum for a very short time duration	Intermittent
Babble	Speech-like, less prominent than a distinct BT	Mostly pervasive
Generator	Low pitch, steady harmonics with constant amplitude	Mostly pervasive
Wind, Breath release	Varying amplitude, no harmonics	Intermittent
School noise	Highly varying amplitude as well as spectral characteristics	Mostly pervasive

Table 2.1: Characteristics of various noise types encountered in the database

2.3 Synthesized Data

For a classifier-based VAD, we require enough data for training. Hand-labeling recordings calls for a lot of manual labor. So the idea here is to annotate a number of clean recordings and then synthesize the noisy data by adding in different types of background noise. If we use 4 noise types, we will finally get four times the amount of original data. A set of 40 relatively clean

audios are selected and are hand labeled for speech regions. Small localized noises, if any, along with a few silence regions are labeled too. All the audios are passed through a program which replaces the localized noise regions with natural silence occurring in the same audio. Thus we get reasonably clean recordings which can be further used for synthesis of noisy data at selected SNR. Detailed statistics of the clean dataset is provided in table 2.2.

The 40 clean recordings were divided into train (20), validation (10) and test (10) dataset. The speakers in the 10 audios used for testing and the noise instances (more about this in the next subsection) chosen to be added in those files are completely different from those used for synthesis of train and validation sets. The train and validation sets have some speakers and noise instances in common. The details can be seen in table 2.3[1].

Data		**Duration (mins)**	**Number of recordings**	**Number of speakers**	**Average speech (%)**	**SNR (dB)**
Natural	Noisy	76	28	21	47	∼17
	Clean	146	40	23	52	∼33
Synthesized Noisy		1168	320 (40 x 8)	23	52	10,20

Table 2.2: Database statistics for the natural and simulated noisy data.

Data	**Average speech (%)**	**Number of audios**	**Duration (minutes)**	**Number of speakers**	**Number of noise instances**
Train	50	20	73	11	8
Validation	54	10	36	10	4
Test	53	10	36	7	2

Table 2.3: Division of the simulated noisy data into train, validation and test sets

2.3.1 Noise data

Of the noise types mentioned in table 2.1, we select four to be added to the clean dataset. Rain is chosen as a representative of stationary noises. Babble and children playing are selected as they are the most prevalent ones and quite challenging too due to their non-stationarity. Wind, which is yet another peculiar and significantly observed noise type, is selected as the fourth type. This type is also representative of breath releases. These noises are added to the 40 clean audios at two different SNRs[2] - 10 dB and 20 dB. These values are inspired by the real-world observed cases for the noise types of interest. 12 different instances for each of these noise types have been used. A different instance of the same noise could be recorded in a different setting, thus giving a variety. For the case of wind noise, different samples constitute the sounds recorded in indoors, outdoors, with different distances from mic, with a slight thunder in between, etc. In

[1]The numbers are as per 40 recordings. For the noisy synthesized dataset (total of 320 recordings), the duration and the number of audios will shoot up to 8 times. The number of noise instances will then be 4 times (1 for each noise)

[2]SNR is evaluated by taking into consideration the energy of clean signal and noise in just the speech regions of the signal. This is possible because we know the speech/non-speech ground truth.

the case of rain, different instances include the audios of rain on rooftop, in open field, heavy rain, light rain etc. Similarly, children playing and babble sounds can vary based on where they have been taken, like babble could be the one recorded in a party, or in a restaurant, or in a marketplace, etc. The sources of these noises are varied and are listed in table 2.4, along with total duration of the collected audios. The total duration varies based on the availability of these audios.

Noise Type	**Number of instances**	**Net Duration (minutes)**	**Source**
Babble	12	15.68	[15], [16], [17]
Children Playing	12	37.21	[15], [16], [18]
Rain	12	329.43	[19], [18]
Wind	12	23.14	[20], [18]

Table 2.4: Noise Data

2.4 Dataset for ASR-based Evaluation of the VAD System

A different dataset comprising of utterances which gave a poor performance on the state-of-the-art ASR system [21], in terms of the phone error rate, were selected. This dataset has a total of 2951 utterances across 43 speakers, having an average duration of 4.5 seconds per utterance. These utterances are extracted from 488 story recordings of 7 different stories. No speaker, nor any story is common with the training dataset. The complete story audio recordings are first passed via a preliminary energy-based VAD[3], where the 4 to 6 minutes audios are chunked into individual story sentences using the video time stamp information. This chunking process makes an assumption that the sentence is read within the time for which it is present on the screen. As VAD plays a role in chunking, a good fraction of low energy non-speech regions has already been discarded. So, this dataset does not have close to 50% non-speech audio, unlike our previous datasets. In chapter 5 we compare the word error rates obtained on this dataset with and without VAD pre-processing and analyze the results in detail.

[3]This VAD is based on the adaptive linear energy-based detection technique discussed in the next chapter, section 3.1.1

Chapter 3

Literature Review

The eventual aim of the project is to correctly recognize the words spoken by the users and provide automated feedback on the reading skills. It may seem like an ASR is a one-stop solution to this problem, but given the school recording environment and our target being the kids from rural areas who are learning English as a second language, our task is filled with challenges due to diversity in speaking accents, pronunciation variability, and the ubiquitous presence of background noise. As previously mentioned, we move on to voice activity detection (VAD) as a preliminary solution to the latter. Voice Activity Detection (VAD) refers to the problem of distinguishing foreground speech segments, could be clean or noisy, from the background noise in an audio stream.

A typical VAD consists of two parts - a feature extractor and a speech/non-speech decision mechanism. The first part extracts relevant speech/noise-specific features which help distinguish the speech frames from the other frames, and the rules set out by the decision mechanism give out the final frame-level decisions. Once these two things have been set out, the next task is to check the reliability and robustness of the proposed VAD algorithm. Depending on the purpose the VAD has been designed to solve, a relevant dataset is used for performing the experiments, and the results are reported using suitable evaluation criteria. In our case, the purpose of VAD is to be a front end for a speech recognition system, but there is very little work on this specific application of VAD [22, 11, 23]. Benchmark algorithms, ones which have been proved to be superior in the literature, are used to further test the applicability of the proposed VAD for the task.

We categorize the review on VAD techniques into 4 sections. Discriminative characteristics of speech in various domains and the different decision techniques are presented in section 3.1 and section 3.2 respectively. Various evaluation criteria are reviewed in section 3.3. Section 3.4 talks in details about the popular benchmark algorithms.

3.1 Features for VAD

Feature extraction is the first step of an automatic VAD system. Most of the features used for such a system are evaluated over a few milliseconds frame. The underlying assumption is that the features are statistically stationary over these few milliseconds period. So the audio signal is divided into overlapping frames, and the decision period is governed by the shift duration. A

good feature is characterized by a reasonable discrimination power between noisy speech and noise. Whether or not the discrimination power is "reasonable" is checked by different metrics designed for the evaluation purpose. [10, 24] are some of the works which have a good compilation combined with a detailed comparison of various VAD features and algorithms proposed over the years.

3.1.1 Energy-based features

Short-time energy is the most common feature, especially in the scenario of a recording via a close-talk microphone. In such cases, foreground speech is expected to have a higher energy than background disturbances captured via a microphone. This feature loses its discriminative power rapidly as the SNR degrades, the soft speech regions are the first ones to be affected. But the energy based features are still popular because of their simplicity in implementation and low computational power. So most of the times, this feature is either used along with other features which deal with the shortcomings of short-time energy [25, 26] or intelligent noise-level dependent adaptive thresholding schemes are implemented. Sakhnov et al. introduce one such example of the latter in [27]. One can also exploit short-time energy in different frequency sub-bands [28]. These sub-band ranges are motivated by the facts that voiced speech has higher energy in low-frequency bands (<2kHz) and the unvoiced and nasal speech regions have higher energy in high-frequency bands (>2kHz) [24]. Some examples of VAD implementation which use energy as one of the features would include [25], which employs multiple features including sub-band energies and [29], which uses short-time energy as a preliminary event detector before further processing and detection-classification. A detailed comparative study of various VAD algorithms which use energy as the primary feature, both in the time domain and the frequency domain, has been discussed in [30].

Adaptive Linear Energy-based Detector

We use the adaptive linear energy based detector (ALED) as one of the features in our proposed VAD approach. Sakhnov et al.[27] have proposed ALED, where the threshold is updated on every detected non-speech frame. As a part of the algorithm, a fixed length (m) buffer of energy of most recently detected m non-speech frames is maintained. The threshold is updated as a weighted average of current threshold and the energy of the most recent frame classified as non-speech. The weight depends on the change in noise variance across time as observed from the ratio $r = \frac{\sigma_{new}}{\sigma_{old}}$. Here, σ_{new} and σ_{old} are the estimated noise variance values calculated using the buffer. A higher value of ratio, r, means that the noise variance changed by a larger amount, i.e. the current noise frame has a higher energy than the previous frames. Hence, more weight is given to the energy of the current frame. The exact value of weight, p, is decided based on the look-up table 3.1. The threshold thus adapts according to the varying noise statistics. The threshold is initialized as an average of the frame-level short time energy for the first 2 seconds, which can be reasonably assumed as non-speech for our task, of the audio.

3.1.2 Spectral-domain features

The speech signal is characterized by specific attributes in the frequency-domain, one of which has been mentioned in the previous paragraph. As the noise is assumed to be additive, most of

Ratio	Value of p
$r \geq 1.25$	0.25
$1.25 > r \geq 1.10$	0.20
$1.10 > r \geq 1.00$	0.15
$1.00 > r$	0.10

Table 3.1: Value of p depending on $r = \frac{\sigma_{new}}{\sigma_{old}}$. Look-up table for algorithm1 (ALED)

the derived features rely on the fact that clean speech spectrum can be derived by subtracting the estimated noise spectrum from the original spectrum of noisy speech, as in equation 3.1 [31].

$$|\hat{X_k}|^2 = |X_k|^2 - |\hat{N_k}|^2 \tag{3.1}$$

where, $\hat{X_k}$, X_k, and $\hat{N_k}$ denote the estimated clean speech power spectrum, original speech power spectrum, and estimated noise power spectrum respectively for the k^{th} frame. Though the noise is additive in both time-domain as well as frequency-domain, it is more desirable to implement such methods in latter as noise and speech are more separable in spectral-domain; there is a higher possibility of remnants of one being found in another if such a separation is attempted in time-domain. This method works by assuming that the noise component is additive and independent from speech in the power spectrum domain. The process of noise estimation can be achieved by using a specific noise training data or by averaging the long-term spectrum of the signal [32, 33]. These methods, though widely used as a part of speech enhancement techniques, they are sometimes included in VAD systems either explicitly as a pre-processing step [34], or implicitly in the feature extraction step itself [32, 35].
Spectral entropy [36] is another spectral-domain measure derived from the speech spectrum. The entropy reflects the flatness of the spectrum and is maximized when all spectral values are equal. For speech, the entropy is low as some frequencies are excited and dominate the spectrum. Whereas for stationary background noise, high entropy is assumed. It is motivated by the observation that the speech regions in a spectrogram are more "organized" than noise regions. This is because speech regions have a formant structure which leads to a specific arrangement of high and low energy regions in frequency domain while this is not the case for most of the noise signals. Another feature closely related to the entropy is the spectral flatness measure [37]. It is based on the ratio between geometric and arithmetic mean of the spectral values.

An appropriate metric to measure the randomness of the signal is Shannon's entropy. So the speech spectrum at each frame is obtained via fast Fourier transform (FFT). The probability density function (pdf) of the spectrum is estimated by normalizing the energy of each frequency bin by all the frequency components, as in equation 3.2. This is calculated individually for each frame. Like energy, entropy can also be computed in sub-bands.

$$p_i = \frac{s(f_i)}{\sum_{k=1}^{N} s(f_k)} \tag{3.2}$$

where $s(f_i)$ is the spectral energy for the frequency component f_i, p_i is the corresponding probability density, and N is the total number of frequency components in FFT. Once we have

the pdf, spectral entropy for each frame can be found out using Shannon's formula as

$$H = -\sum_{k=1}^{N} p_k log(p_k) \tag{3.3}$$

Taking the concept of entropy a step further, relative spectral entropy (RSE) [38] is defined as entropy with respect to the mean spectrum. This is helpful in the case where the audio recording has a persistent noise source present. RSE is the KL divergence between the current spectrum (p_k) and the local mean spectrum (m_k) computed over some number of neighboring frames.

$$RSE = -\sum_{k=1}^{N} p_k log \frac{p_k}{m_k} \tag{3.4}$$

Using entropy alone might fail when the present noise has an "organized" spectrum too, for example, bell, generator, music. So, [26] proposes a new feature by combining entropy with energy. The motivation is that the combined feature will help alleviate the shortcomings of both the individual features. In the presence of non-stationary sounds like children playing, entropy would be more reliable than energy. Energy can help for the cases where entropy fails and the noise has a low amplitude.

3.1.3 Harmonicity-based features

Speech is modeled by a voiced or unvoiced excitation signal that is spectrally shaped by the vocal tract. Pitch and harmonicity are the features that target at the voiced excitation. For voiced regions, vibration of vocal cords produces a harmonically rich sound with pitch lying between 50 and 250 Hz [39]. Most of the phones exhibit this property, making this characteristic is very peculiar to speech. However, harmonicity-based features fail for unvoiced speech regions like fricatives [40]. Also using just this feature alone, harmonic noises like generator will get misclassified as speech. In [41], the harmonicity feature has been combined with energy measure as a compromise in cases where background noise with strong harmonic structure has energy lower than speech. A number of features, of varying complexity, that are designed to capture the harmonic structure of speech are investigated in [42].

Zero-crossing rate is one of the simplest features. It is typically higher for unvoiced speech than for voiced speech due to the dominance of high-frequency energy in the unvoiced consonants. Another measure known as zero frequency filtering has been proposed for epoch detection in speech signals [1]. The zero frequency filtered (ZFF) signal is further used in [43] for the separation of foreground speech from noise. Most harmonicity-based techniques would suffer in environments containing cross-talk speech, such as babble noise or overlapped speaking. The interference of other speakers causes the spectral frames of voiced speech to possibly contain more than one fundamental frequencies, therefore degrading the harmonicity. In [44] a new feature based on auto-correlation is proposed to detect the presence of a disturbing interferer's speech along with the foreground speech. However, the proposed technique was only studied in the case of overlapping speech from two speakers; no experimental results for noisy environments such as babble noise were reported.

Zero Frequency Filtered Signal

We use a feature derived from ZFF output [1] of the given speech signal. The presence of pitch is the outcome of impulse train from source. So, when these impulses are passed through the vocal tract filter, small discontinuities are introduced in its impulse response. The locations of these discontinuities are exactly the positions of glottal epochs. Next, if we pass this speech signal with inherent discontinuities via a resonator, we will get a single frequency in the output, but the discontinuities will remain. For example, in figure 3.1 where the input is the simply the impulses instead of the speech signal, the instantaneous frequency of the resonator output is at 0.39 (radians equivalent of 500 Hz), which is same as the resonator frequency, except for the locations of impulses. Similarly, when we pass a speech signal, which is nothing but impulses modulated by the vocal tract filter, through the resonator, the output will have some discontinuities at the epoch locations; as shown in figure 3.2. A differenced EGG signal has also been shown for the sake of comparison.

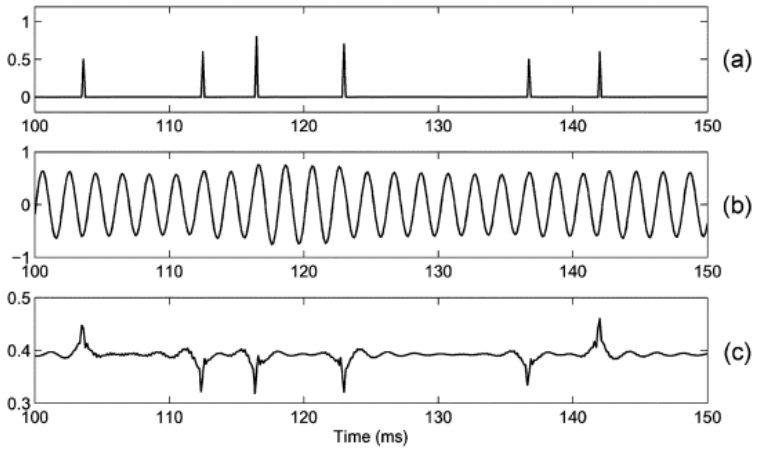

Figure 3.1: Aperiodic sequence of impulses filtered through a 500-Hz resonator. (a) sequence of impulses with arbitrary strengths, (b) resonator output, (c) instantaneous frequency of the filtered output [1]

Ideally the resonator frequency could be any number, but in the case of speech signals, the fluctuations due to the time-varying vocal-tract system could also get captured. So, it is difficult to extract the instants of excitation from the instantaneous frequency computed around an arbitrary center frequency. A zero frequency resonator system is chosen so that the characteristics of the time-varying vocal-tract system will not affect the characteristics of the discontinuities in the resonator filter output. The ZFF signal is obtained by passing the original speech signal through a cascade of two (in order to reduce the effects of all high-frequency resonances) 0-Hz resonators followed by mean subtraction [45]. As the 0-Hz resonator is equivalent to the double integration of the signal, the outcome is an exponentially growing or decaying signal, making it difficult to capture discontinuities caused by pitch impulses. So, mean subtraction needs to be performed at the frame level, where average DC value of a frame is subtracted from all the samples of that frame. The series of operations performed are mentioned in equations 3.5,3.6,3.7.

1. Difference the speech signal $s[n]$ in order to remove and time-varying low-frequency bias in the signal

$$x[n] = s[n] - s[n-1] \tag{3.5}$$

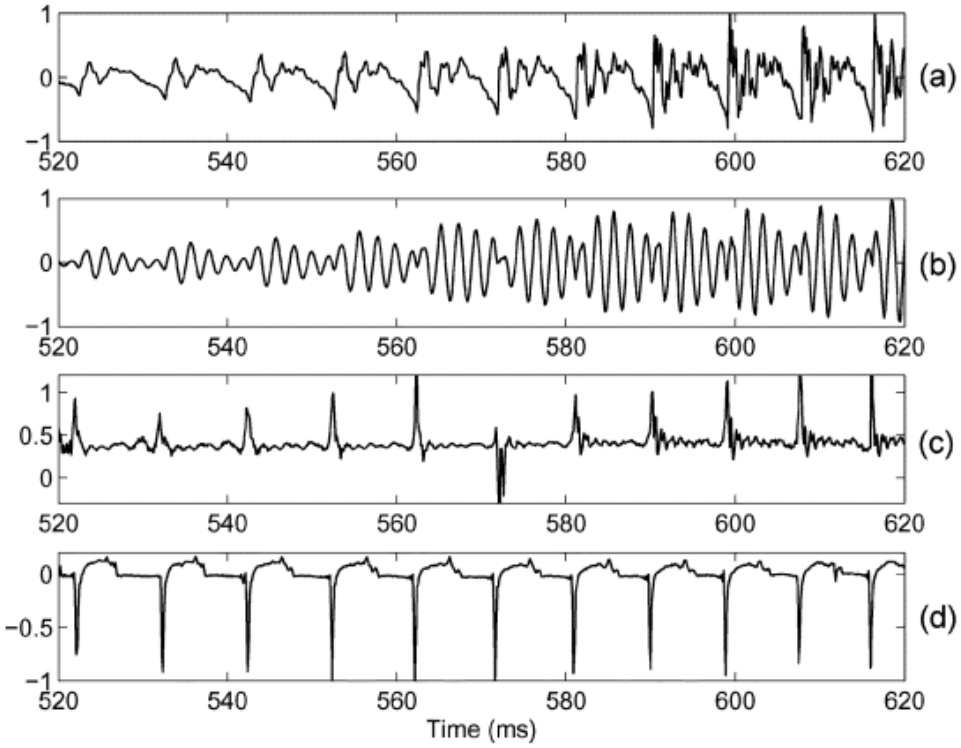

Figure 3.2: A 100-ms segment of (a) speech waveform, (b) output of the resonator at 500 Hz, (c) instantaneous frequency of the filtered output, and (d) differenced EGG signal [1]

2. Pass the differenced speech signal twice through an ideal resonator at zero frequency

$$y_1[n] = 2y_1[n-1] - y_1[n-2] + x[n] y_2[n] = 2y_2[n-1] - y_2[n-2] + y_1[n] \tag{3.6}$$

3. Remove the trend in $y_2[n]$ by subtracting the average over 10 ms at each sample

$$y[n] = y_2[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_2[n+m] \tag{3.7}$$

This operation is repeated thrice in order to completely get rid of the increasing/decreasing trend [45]

3.1.4 Long-term features

Long-term speech temporal information [35, 32, 46, 47] has been the focus of research in the speech community recently. A person, on an average, has a speaking rate of approximately 10-15 phonemes per second [48]. Each of these phones exhibits different spectral characteristics and thus the speech statistics to vary greatly over time. Based on the assumption that most of the day-to-day background noises have a degree of variation lower than that present in normal speech, analysis of speech over a longer window is done for speech-noise separation. [34] finds the long-term spectral envelope, from which a measure is derived from the subband SNRs. [35] proposed the long-term dynamic feature for VAD, derived from fusing the delta cepstrum of the neighboring frames. Sample variance of the long-term sub-band entropies has been calculated in [46], showing a significant improvement in both stationary and non-stationary noise conditions. VAD based on long-term spectral flatness measure has been proposed in [49] and is claimed to

outperform [46] on challenging noise types such as babble.

The limitation on the rate of speech is due to the physical constraints on the speed of the articulatory movements. Because of this, there is an inherent low-frequency modulation in a highly varying speech signal. In [50] it has been shown that the modulation frequencies below 16 Hz contribute to speech intelligibility significantly. Long-term modulation features based on these facts have been proposed in [51, 52]. Amplitude modulation spectrogram (AMS) is one such popular feature. It represents the decomposition of the signal along acoustic frequency, modulation frequency and time, leading to a very high dimensional feature vector. For instance, in [51], the signal is divided into 17 acoustic frequency bands and 29 modulation frequency bands are evaluated for each of these 17 bands; this leads to a 493-dimensional feature every time step. Methods have been proposed for dimensionality reduction of the AMS [53, 54]. In [3], an optimum range for speech frequency and modulation frequency has been analyzed, thus giving a single-valued feature for each shift frame. Modulation feature has been combined with harmonicity and the resultant VAD is used as a pre-processing unit before ASR, as in our task, in [22].

3.1.5 Miscellaneous features

Exploiting the harmonic structure in different frequency bands could be more beneficial depending on the noise type. Hearing research also suggests that the decomposition of the periodic and aperiodic components plays an important role in the human auditory system. Motivated by this, a novel feature, periodic-to-aperiodic component, has been proposed in [55]. Mel-frequency cepstral coefficients (MFCCs), inspired by human hearing, are another set of widely used features [54, 56] in supervised speech/non-speech classification tasks. A gamma tone filtering based feature, inspired by the motion of basilar membrane and motivated by the single frequency filtering [57] technique, and entropy together have been used in [58]. Biologically inspired features related to the computational auditory scene analysis (CASA) research have been used for speech/non-speech classification [59]. It is inspired by the process by which the auditory system separates the individual sounds in natural world situations, in which these sounds are usually interleaved and overlapped in time and their components interleaved and overlapped in frequency [60].

3.2 Decision Mechanism

Once we have the extracted features from the original signal, next step is to categorize the frames into speech or non-speech, which includes construction of a set of decision boundaries that partition the feature space into these two classes. The simplest and most intuitive way to generate decision boundaries is via thresholding, and these thresholds are set either empirically for the complete recording, depending on the feature values [2, 57], or are adapted according to the recording by updating at certain frames [25, 27]. The latter helps the cases where non-stationary background noises are present. Some works employ multiple thresholds to detect both activation and deactivation transitions[26].

Simple thresholding works in the cases where the feature space is linearly separable. But in the case of noise corrupted data, linear decision boundaries might no longer be able to robustly classify the two regions. In such cases, non-linear decision boundaries are used, such as those

based on statistical models and machine learning approaches. A statistical model-based approach derived from the generalized likelihood ratio test was first introduced in [61]. Since then, different models and formulations of hypothesis tests have also been proposed [62, 63, 64, 65].

Recently, various machine learning approaches have been implemented as decision rules for VAD. Support vector machine (SVM) has been one of traditionally used classifiers [66, 67, 68]. Maximum marginal clustering, neural networks, linear discriminant analysis (LDA), boosting algorithms, genetic algorithms are a few of the other widely used classifiers. Maximum marginal clustering, being an unsupervised approach, requires no training labels and is observed to perform better than SVM [69]. In [70] a multi-layer perceptron classifier has been trained on 4 features based on different characteristics of speech. In [71], a 3-layer feedforward neural network has been trained on MFCC features. LDA can be used as both dimensionality reduction technique as well as a method to fuse multiple features. Boosting is used to combine a set of weak classifiers to make a stronger one [72]. Adaboost is applied to both long and short duration features in [73]. Genetic Algorithms are randomized search and optimization techniques guided by the principles of evolution and natural genetics [74]; here, zero crossing rate and a new feature extracted from signal envelope are used to train the genetic algorithm. As per [24], most of the above-mentioned machine learning based approaches, though show improved VAD results, have not been compared extensively with modern VAD techniques. But recently a single-hidden-layer deep belief network based approach was proposed in [75]; it is evaluated against 11 different established VAD algorithms on 7 noise types, including babble. Along with accuracy, the work also compares the time complexity of all the VAD approaches.

3.3 Evaluation Criteria

The choice of an evaluation metric depends on the end goal of the application. For instance, in a speech codec application, as one cannot afford to lose any speech segments, a high detection rate of speech frames is desired. In that case, minimizing the wrongly detected speech frames (false alarms) is a lower priority task. The final aim of our task is to improve the speech recognition accuracy by removing the non-speech regions from the recordings using a VAD as a pre-processing step. So the ideal way to evaluate this VAD would be to run ASR experiments with and without any pre-processing and compare the phone/word error rate measure to check whether such pre-processing helps or not and by what amount. But because of the time complexity of running an ASR engine, it would be desirable to have a VAD performance metric which is directly correlated with the ASR accuracy on the VAD pre-processed data.

3.3.1 Frame-level

Traditionally, frame-level accuracy [71], as in equation 3.8, has been the most basic criteria for evaluating VAD performance.

$$\text{Accuracy} = \frac{\text{Total number of correctly detected frames}}{\text{Total number of frames}} \tag{3.8}$$

$$\tag{3.9}$$

The more detailed frame-level evaluation metrics of hit rate (HR) and false alarm rate (FA) are more widely used. As it is clear from the equations 3.10, ideally a high HR and a low FA is

desired.

$$\text{HR} = \frac{\text{Number of speech frames detected as speech}}{\text{Total number of speech frames in ground truth}} \tag{3.10}$$

$$\text{FA} = \frac{\text{Number of noise frames detected as speech}}{\text{Total number of noise frames in ground truth}} \tag{3.11}$$

These metrics give us extra information as compared to the accuracy alone, which is not so reliable unless we have a good distribution of both speech and non-speech segments. For instance, if the recording has speech content for 90% duration, then even if the VAD outputs all ones, accuracy will be quite high at 0.9. Accuracy won't be able to capture the fact that our VAD completely fails to detect non-speech segments. So, we cannot determine whether the algorithm is good at detecting speech segments or at detecting the non-speech ones by using just accuracy. Whereas, in the case of HR and FA we can clearly identify which of speech/non-speech regions are causing a problem. For the previously mentioned example, HR would be 1 but the FA will also be 1, thus hinting that the VAD cannot detect non-speech segments at all.

The issue with HR and FA is that there is always a trade-off when we change the threshold in our decision mechanism. For a simple example of using short-time energy (subsection 3.1.1) as a feature, we need to threshold it to get to a binary classification. Now if the threshold is set to be too low, we will surely be successful in correctly detecting all of the speech frames but many non-speech frames will be wrongly detected as speech. Thus leading to a high HR at the cost of high FA. Similarly, setting the threshold too high will lead to a low FA but at a cost of low HR. So, another metric based on HR and FA, receiver operating characteristic (ROC) [32, 76], is used especially for the cases where two or more VAD algorithms are to be compared. ROC is a plot of false positive rate (FA) vs true positive rate (HR). If one curve is completely above another then the VAD to which the former curve corresponds to is better. Though better than accuracy, HR and FA still do not say much about the specific locations in the speech segment where the errors occur.

First introduced in 1989 [77], a metric which better represents the errors is being used for evaluation purposes more recently [49, 58]. More advanced than HR and FA, it accounts for errors happening at the speech onset, within the speech segment and at the speech offset. In short, the accuracy is divided into the following 4 metrics

- Front-end clipping (FEC):
  Clipping due to speech being misclassified as noise in passing from noise to speech activity
- Mid-speech clipping (MSC):
  Clipping due to speech being misclassified as noise during a speech region
- Carry-over (OVER):
  Noise interpreted as speech in passing from speech activity to noise
- Noise detected as speech (NDS):
  Noise interpreted as speech within silence/noise region

Figure 3.3 clearly depicts which respective frames will contribute to these 4 metrics using an example ground truth and VAD output labels. The metrics are evaluated as the number of

frames belonging to respective class divided by the total number of frames. From the figure, we can also observe that NDS and OVER together make up for all the false alarms and FEC and MSC together constitute all of the misdetections. While frames are the most basic entity, we are effectively considering errors in different positions of speech segments, so these metrics could be considered as an intermediate between frame level and segment level evaluation criteria.

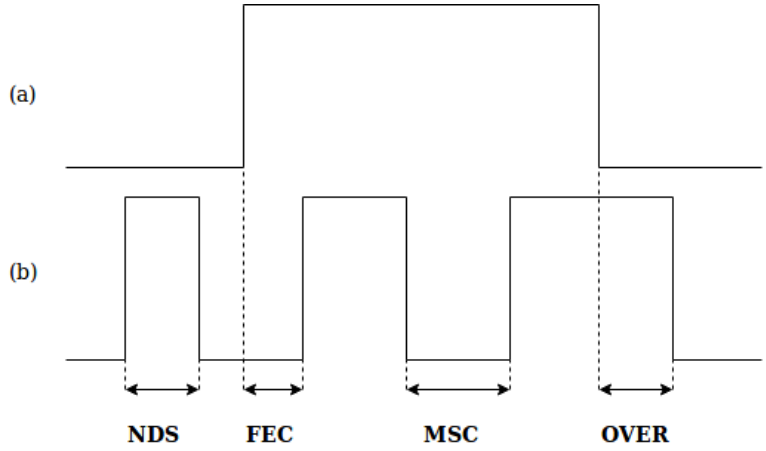

Figure 3.3: Example explaining the frame-level evaluation metrics, where high denotes speech and low denotes non-speech. (a) Ground truth labels; (b) VAD output labels

3.3.2 Segment-level

One issue with the above-mentioned metrics is that it is difficult to decide the relative importance of these four metrics as we do not know which one affects the ASR accuracy more than the rest. Moreover, there have been no experiments in the literature to show that these criteria are actually correlated with ASR accuracy [78]. In [11], where an experimental analysis of the relationship between VAD and ASR accuracy has been done, it is shown that the boundary effects, which directly influence start and end of the sentence are known to be very significant for ASR tasks. So, inspired by the preceding discussion and the fact that the number of segments has a crucial influence on ASR accuracy [11], a new combined segment-level evaluation metric has been proposed in [78]. It has been shown to give a positive correlation with ASR performance. The proposed combined metric is the harmonic mean of the following four values

1. Start Boundary Accuracy (SBA)
   This is the quantification of the overlap region of the ground truth speech onset and VAD speech onset, within a certain tolerance. In the figure 3.4, assume that the VAD output position at speech to non-speech transition is fixed and at speech onset, it varies between points (a) and (b). In such a case the ASR accuracy would not vary much (within a fixed tolerance of variation), so we need the metric to not penalize when the actual speech onset happens a bit after the predicted one. Now consider the points (b) and (c), as the boundary shifts from (b) towards (c), ASR accuracy will go on degrading and so is expected from the error metric. So the SBA is proportional to the amount of overlap between ground truth and the VAD output in the region within (b) and (c) at the speech onset. Same could also be seen analytically from equation 3.12.

2. End Boundary Accuracy (EBA)

Exactly the same explanation as above holds for EBA, with speech onset replaced with speech offset. The ASR accuracy will reduce only when the end boundary in figure 3.4 starts shifting from point (b) towards (c), and so should EBA. So the EBA is proportional to the amount of overlap between ground truth and the VAD output in the region within (b) and (c) at the speech offset. Equation 3.14 for end boundary accuracy, similar to the previous one, is self-explanatory.

$$J_s^r = \frac{\sum_{i\in[s_r-L,s_r+L]} f(i-s_r)\delta(x_i, x_i^{ref})}{\sum_{i\in[s_r-L,s_r+L]} f(i-s_r)} \tag{3.12}$$

$$SBA = \frac{\sum_{r=1}^{R} J_s^r}{R} \tag{3.13}$$

$$J_e^r = \frac{\sum_{i\in[e_r-L,e_r+L]} f(e_r-i)\delta(x_i, x_i^{ref})}{\sum_{i\in[e_r-L,e_r+L]} f(i-s_r)} \tag{3.14}$$

$$EBA = \frac{\sum_{r=1}^{R} J_e^r}{R} \tag{3.15}$$

where,
N: number of speech segments in the ground truth,
L: tolerance duration around the boundary (this might vary according to the length of the speech segment),
$f(x)$: weighing function such that heavier weight is given to the frames within the reference speech segment,
s_r: frame number at the start boundary of r^{th} reference speech segment,
e_r: frame number at the end boundary of r^{th} reference speech segment,
R: Total number of speech segments,
x_i: class of the i^{th} frame in hypothesis,
x_i^{ref}: class of the i^{th} frame in ground truth

3. Border Precision (BP)
   This accounts for the fact that higher number of segments in hypothesis than in ground truth degrades the VAD performance. Let N_{gt} be the actual number of segments and let N_{vad} be the number of segments in the VAD output. Then BP is defined as in equation 3.16. So higher the number of segments, lower will be the ASR accuracy and so the BP. This metric in a way decides how reliable are the previous two metrics depending on the number of speech segments. For example in the example shown in figure 3.5, the SBA and EBA for the VAD output is quite high but the ASR performance will not be too good because of too many missing speech frames in between. These missing speech frames lead to an increase in the number of segments and thus will degrade the BP.

$$\text{BP} = \frac{N_{gt}}{N_{vad}} \frac{EBA + SBA}{2} \tag{3.16}$$

4. Accuracy (ACC)
   This is the overall accuracy, same as equation 3.8

The harmonic mean is used as it tends to mitigate the impact of large quantities and aggravates the impact of small ones.

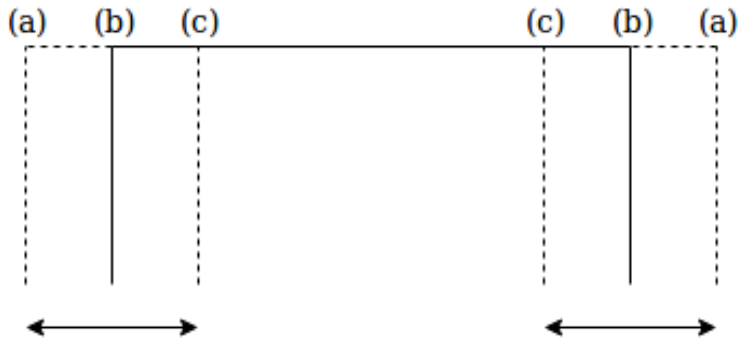

Figure 3.4: Example for motivating the SBA and EBA, where high denotes speech and low denotes non-speech. Solid line: Ground truth label; Dashed line: VAD output label variation

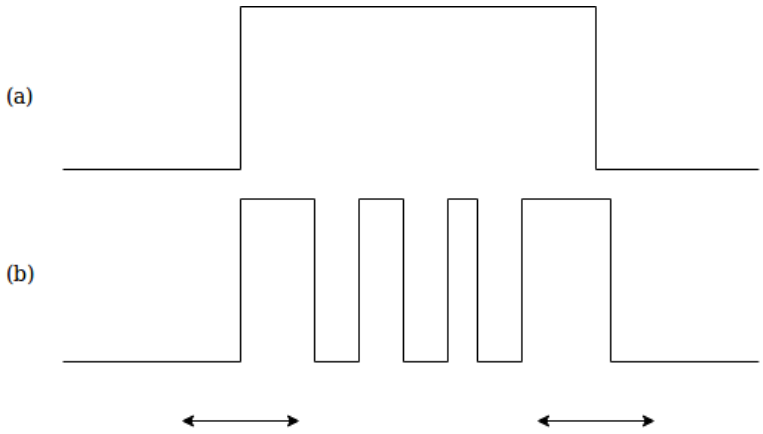

Figure 3.5: Example for motivating BP, where high denotes speech and low denotes non-speech. (a) Ground truth labels; (b) VAD output labels

Thus, the evaluation metric used is based on the task for which VAD needs to be evaluated. Deciding a reliable benchmark algorithm in order to validate the proposed VAD system is the next important task. So, we look at the different benchmark VAD algorithms that have been established and used over the past few years in the next section 3.4.

3.4 Benchmark Algorithms

Deciding a benchmark is another crucial step in developing a VAD algorithm. Standard speech codec algorithms [2, 25] have an inbuilt VAD which is considered as standardized and has been widely used as a baseline for various VAD applications. The long-term features mentioned in section 3.1.4 have also been quite popular lately [75, 49, 79]. The speech codec algorithms' VADs - Adaptive Multi-rate VAD2 (AMR2) and ITU-T G.729 (G729), have their C implementations available too. Sohn's VAD [76] has also been used to benchmark the statistical VAD models. For benchmarking our system, we use AMR2 and single frequency filtering [57] based VADs. G729 uses sub-band energy, zero-crossing rate, and spectral based features and the decision mechanism is designed as an adaptive threshold. The results by G.729 were found to be clearly underperforming as compared to AMR2, as also reported by Beritelli, et al. in [80], AMR2 was chosen as a benchmark.

3.4.1 Adaptive Multi-Rate

AMR2 decisions are based on simple energy-based features which are derived from signal power extracted in 16 frequency sub-bands using filter banks. It adapts itself for different degradation by estimating the SNR and is seen to give good performance over different SNR situations. Figure 3.6 provides a block diagram of different steps involved in the AMR2 algorithm.

This VAD divides the 20-ms frames into two subframes of 10 ms and calculates the channel power, the voice metrics (voice-quality measure as a function of SNR), and the estimated noise power for each of them. The spectral deviation estimate helps against the erroneous background noise update. The update decision is also based on the long-term prediction gain which is derived using pitch. It is derived as the spectral deviation of power spectrum of the current frame and the average long-term power spectrum estimate. The estimated SNR is quantized between 0 to 19 in 3 dB steps. The decision is made by comparing the voice metrics with a threshold that varies according to the estimated SNR. A hangover scheme is incorporated as well. The frame is classified as speech if at least one of the two consecutive frames could be classified as a speech frame. The codes were obtained from the author's C implementations [81].

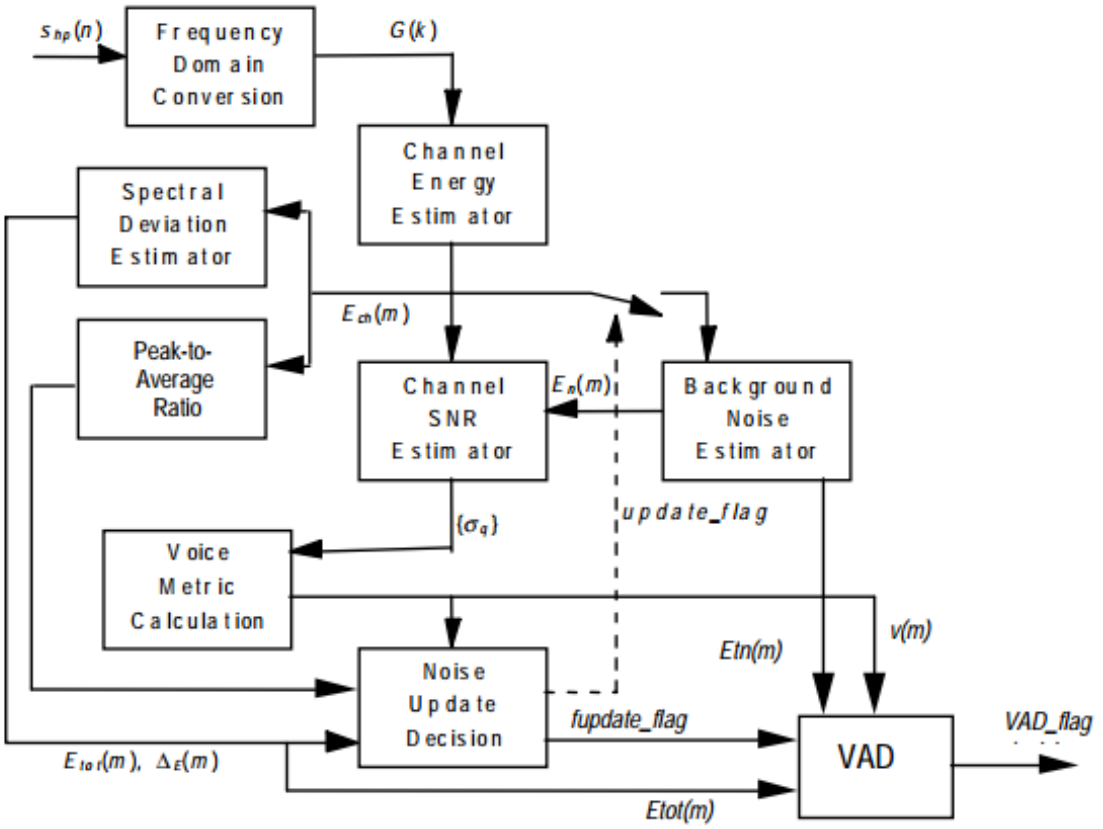

Figure 3.6: Block diagram for AMR2 algorithm [2]

The AMR2 approach neither uses any of the long-term characteristics nor does it exploit any features involving the formant structure or the harmonicity. Though the long-term power estimate and the pitch are indirectly used for the background noise update, it would be challenging to attain a good performance on non-stationary noise types. The decision smoothing criteria is poor as it is based on just 2 frames, i.e 20 milliseconds of speech.

3.4.2 Single Frequency Filtering

Proposed in 2015, [57] exploits the fact that the variance of the spectral intensity across frequency is higher for speech and lower for many types of noise. The key aspect in their algorithm is the high spectral and temporal resolution of the amplitude envelope obtained at each of the 185 frequencies, equally spaced between 300 Hz to 4000 Hz. Due to good resolution, regions in the time-frequency space can be found where SNR is high. The weights for each envelope (the estimated noise floor) is estimated by taking the mean of lower 20% values of the envelope at each frequency; this is based on the assumption that there is at least 20% of non-speech in the recording. Thus we get the noise-compensated envelopes by dividing the envelope at each frequency by the estimated noise floor. As the noise components are de-weighted, the mean of the square of component envelopes across all the frequency components at a time instant is expected to be higher for speech frames. Same is the case for standard deviation, because of the formant structure of speech. Thus, the measure derived from mean and standard deviation across all the frequencies of these noise-compensated envelopes serves as a feature per 10 milliseconds frame. So, although no assumption regarding the noise type or about the nonspeech beginning is made, while deciding the thresholding criteria there is an underlying assumption of at least 20% of the audio having non-speech frames. The threshold, decided using the feature values of the whole recording, is fixed for that particular audio signal. The decision at a sampling instance is made by considering a window size decided on the basis of the dynamic range, as defined in [57], of the complete audio signal. The idea is that a discriminative feature will have a higher dynamic range and a smaller window size should perform well; while if the noise is such that the proposed feature is not discriminative (implied by a lower dynamic range) then a higher window size is considered. The temporal smoothing criteria is also dependent on the dynamic range of the feature. If the dynamic range is high, the decisions are assumed to be more reliable and the smoothing is done over a longer window.

Four types of degradation are considered by the authors - i) due to noise, ii) due to telephone channel, iii) due to cell phone channel, iv) due to distant speech. The results are evaluated with respect to AMR2 method. The dynamic ranges for different noise-SNR pairs have been provided in the paper and it can be observed that the individual performances are proportional to those values (implied discriminative power). Of the noises of our interest, the proposed feature does not give its best performance on babble noise. This could be because spectrally babble is not so different from speech; the spectral power is spread across (non-uniformly) a wide range of frequencies for babble too. Though it is recent, this method has already started being used as a benchmark in a few other VAD literature [58].

Chapter 4

Proposed Approach

4.1 Features

At times it could be possible that a feature which can give a good discrimination between speech and silence, does not perform so well on noise and noisy speech classification. A simple example could be using harmonicity as a feature - speech regions will definitely have higher harmonicity than non-speech regions in a clean recording environment. But as the SNR starts degrading, the noise could deteriorate the harmonicity of speech region, or if the noise is harmonic like a generator sound, then the non-speech regions have an increased harmonicity, unlike the clean speech case. So the choice of features is motivated by the noise types and the SNR to be dealt with as well. Next, we describe and motivate the features used by our proposed method. The final decisions, both by the benchmark and proposed algorithms, are made on a 10ms frame length. So, in all the cases shift is fixed at 10 milliseconds, irrespective of the duration on which the feature is evaluated.

4.1.1 Short-time energy

Short-time energy is evaluated as a root of squared sum of values of 20 milliseconds long hamming windowed signal. This is the most basic feature of all. Another feature we use is derived from the ALED algorithm (described in detail in section 3.1.1. The short-time energy is adjusted by the adaptive threshold of the ALED algorithm, i.e.

$$E_{aled}(i) = E_{ste}(i) - k * E_{th}(i) \tag{4.1}$$

where, E_{aled} is the value of the derived feature, E_{ste} is the short-time energy, E_{th} is the updated threshold value, all at the i^{th} frame, and k is the threshold factor. This feature is introduced to deal with relatively stationary noises like rain, generator. It might also work on babble scenario if the amplitude is more or less constant. This can also be inferred from figure 4.1, where it is observed to perform the best for rain noise and degrades the most for wind noise. For rain, the non-speech portions are almost zeros as if the adaptive threshold perfectively tracked the rain noise level. The performance for babble is decent too, while for children playing case it is not too reliable with clearly visible random spurts in between (60-62 seconds for example).

However, a modification was made in the original ALED algorithm. In the original algorithm, the threshold does not change on detection of a speech frame. Hence threshold goes on decreasing as more number of low energy noise frames add to the noise frame buffer. Instances were

observed where the threshold reached a value lower than most of/all the subsequent non-speech frames and then a large chunk of the recording ended up getting misclassified as speech. As a corrective measure for this, we modified the original algorithm to keep a tab on the length of segment getting classified as speech. For the story-reading task, the continuous speech was observed to be not more than 2.5 seconds; so if the detected speech segment becomes longer than this we reset the threshold, and resume the algorithm from the beginning of the on-going speech segment. Algorithm 1 provides the details of modified ALED.

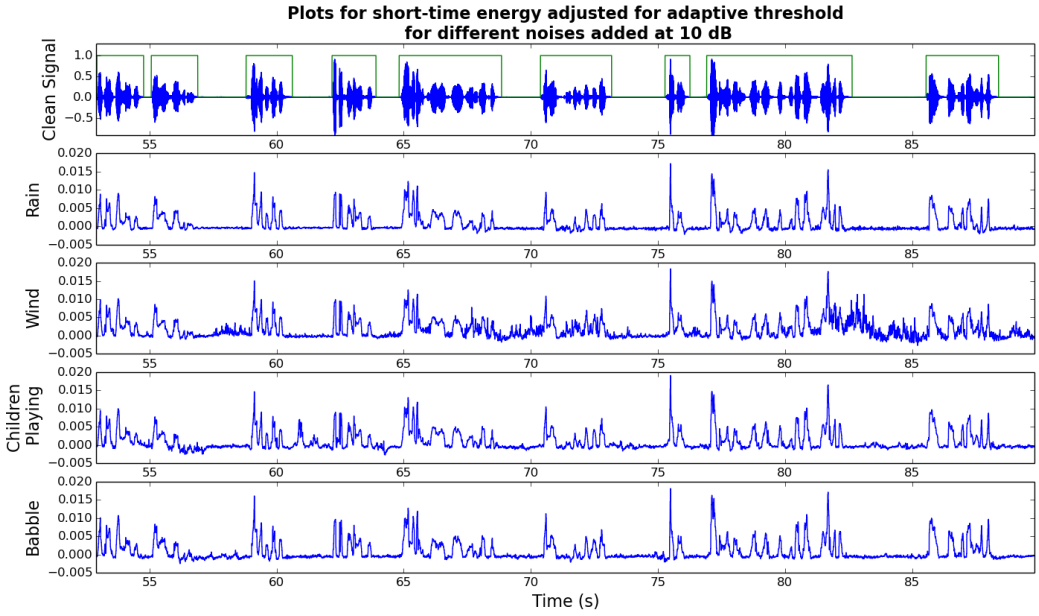

Figure 4.1: An example for short-time energy adjusted for the adaptive threshold extracted for different noise types. Green signal in the first plot indicates the groundtruth labels.

4.1.2 Harmonicity measure

Another conventional feature widely used in VAD is harmonicity. We discuss the motivation of introducing zero frequency filtering (ZFF) in section 3.1.3. In [43], the normalized first order correlation coefficient of the ZFF signal is used as a harmonicity measure for speech/non-speech detection. A plot of first order correlation coefficient is shown in figure 4.2. As expected, it is found to be high in the voiced speech regions of foreground speech as compared to other regions. The main motivation of introducing this feature is to deal with completely non-harmonic noises like wind and breath release. The performance of zffs degrades a lot in the presence of harmonic noises like generator. As is visible from plots in figure 4.3, in the case of wind noise this feature shows a clear demarcation between noise and speech. It performs quite well on rain noise, as rain is non-harmonic too. Next, as the harmonic components in noise increase, this feature fails to discriminate between speech and noise; as is visible in case of children playing and babble.

4.1.3 Spectral entropy

Spectral entropy, discussed in detail in section 3.1.2, exploits the organized formant structure of speech regions. Before evaluating the probability density function, the spectrum is divided into

Algorithm 1 Modified ALED Algorithm

Initialize k, $speechLengthThreshold$, $E_{th,0}$, $winSize$, $bufferSize$, $fixedLengthBuffer$, j, $prevVar$
while $j < numberOfFrames$ **do**
  if $E_j > kE_{th,j}$ **then**
    $decision[j] \leftarrow 1$
    $length \leftarrow length + 1$
    if $length = 1$ **then**
      $tap \leftarrow j - 1$
    if $length > speechLengthThreshold$ **then**
      $j \leftarrow tap$ #Pointer for threshold reset check
      $E_{th} \leftarrow hardThreshold$ #Resetting Threshold
  else
    $length \leftarrow 0$
    $decision[j] \leftarrow 1$
    update $fixedLengthBuffer$
    $var \leftarrow$ variance *(buffer)*
    $ratio \leftarrow \frac{var}{prevVar}$
    $p \leftarrow$ look-up-table *(ratio)*
    $E_{th,j+1} \leftarrow (1-p).E_{th,j} + p.E_{silence,j}$
    $prevVar \leftarrow var$
    if $E_{th,j+1} > hardThreshold$ **then**
      $hardThreshold \leftarrow E_{th,j+1}$
  $j \leftarrow j + 1$

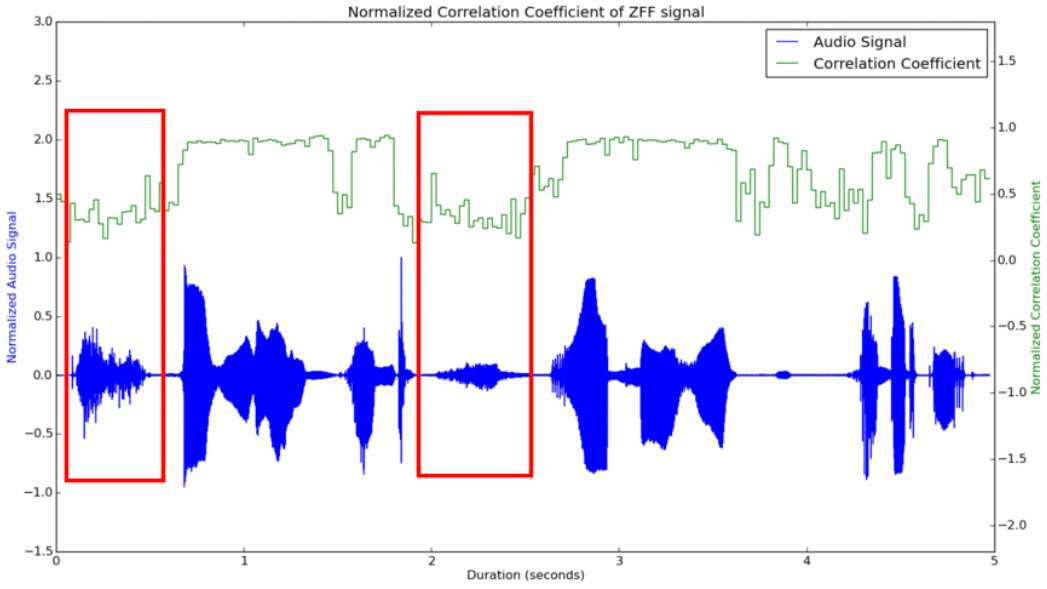

Figure 4.2: Normalized First Order Auto-correlation Coefficient of ZFF signal. Red boxes denote the regions of BR (non-harmonic), thus showing a drop in the corresponding correlation plot.

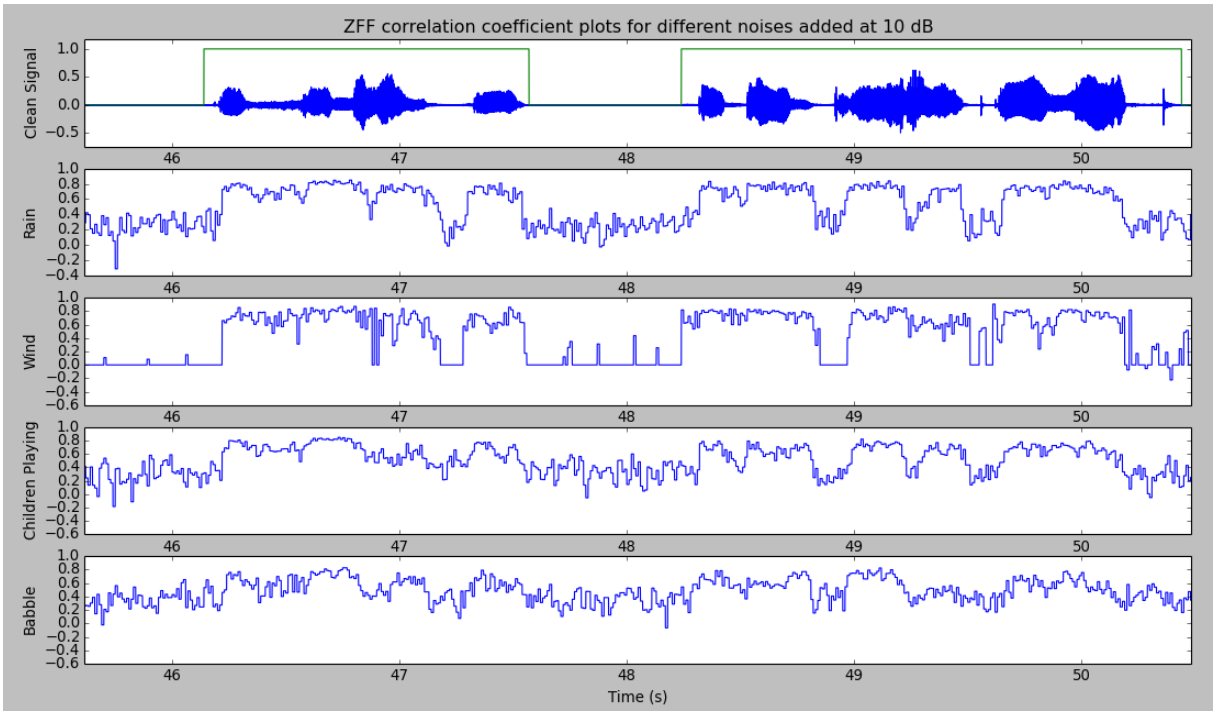

Figure 4.3: First order correlation coefficient of ZFF extracted for different noise types. Green signal in the first plot indicates the groundtruth labels.

three sub-bands. The range for these sub-bands is decided based on the location of formants. The first two sub-bands ranges are fixed at 250 Hz to 850 Hz and 1000 Hz to 2500 Hz respectively [82]. The third sub-band is fixed at 4000-6000 Hz. The presence of strong wind in the signal affects this feature very badly, especially the lower frequency bands. This is because the wind energy is concentrated in lower frequency bands too and so, better discrimination is observed in the 4000Hz to 6000Hz band. This phenomenon can be clearly observed in figure 4.4. The third sub-band has the best performance of all the three sub-bands. In fact, in the lower sub-bands, speech and non-speech are completely indistinguishable. This is because of the prominent presence of low-frequency components in wind noise. On the other hand, in case of rain noise (an example is shown in figure 4.5), lower energy sub-bands are much better. This is because energy is almost uniformly spread across all the frequency components in case of rain. This feature fails miserably in case of babble noise, as speech and non-speech have same characteristics as far as formant structures are concerned.

4.1.4 Zero-crossing rate

Zero-crossing rate (ZCR), yet another very simplistic feature, is typically higher for unvoiced speech than for voiced speech. From the plots in figure 4.6 it is clear that this feature performs well in case of wind noise. As the wind is a low-frequency noise type, it does not degrade the ZCR of noisy speech regions much. ZCR is known to be robust to babble noise as well [83]. Children playing noise has high frequency sounds like shouts and screams in between, because of which ZCR is seen getting affected in some parts as compared to that in case of babble. This feature performs the worst in rain noise, which could mainly be because of the presence of high-frequency components in rain. This not only leads to a higher ZCR in noise regions but the presence of too many frequencies in the background masks the effect which speech should

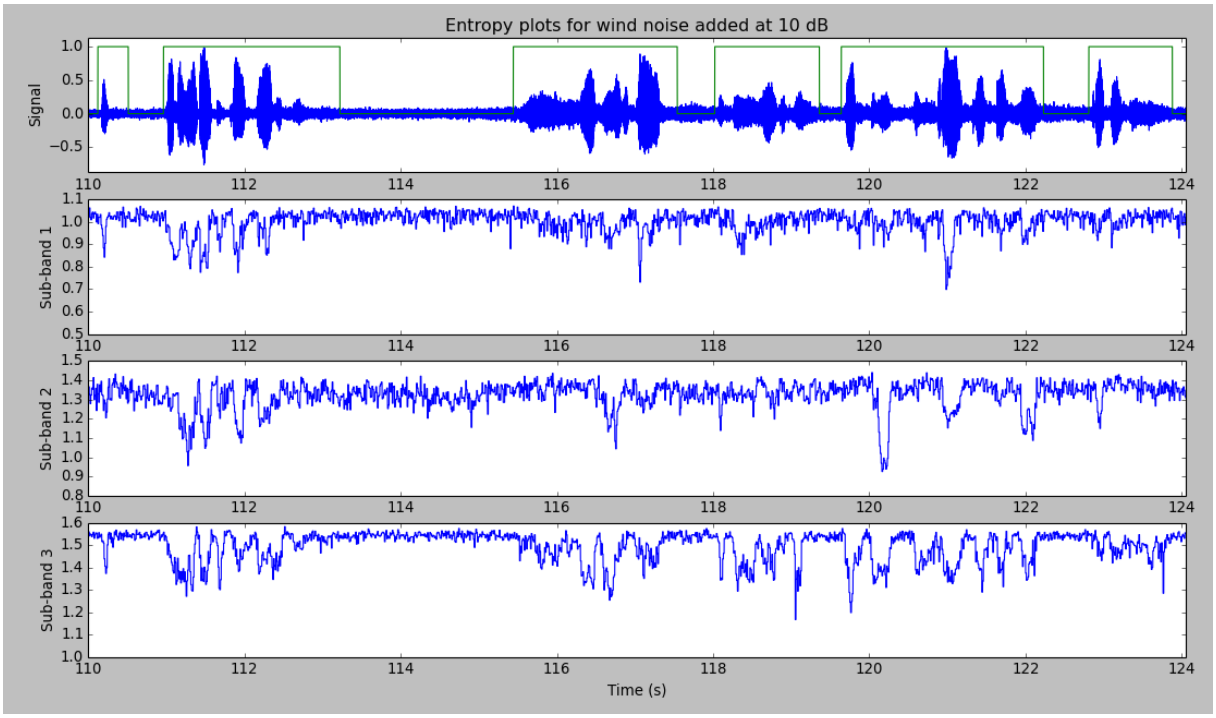

Figure 4.4: Example of entropy feature extracted for the case of wind noise. Green signal in the first plot indicates the groundtruth labels.

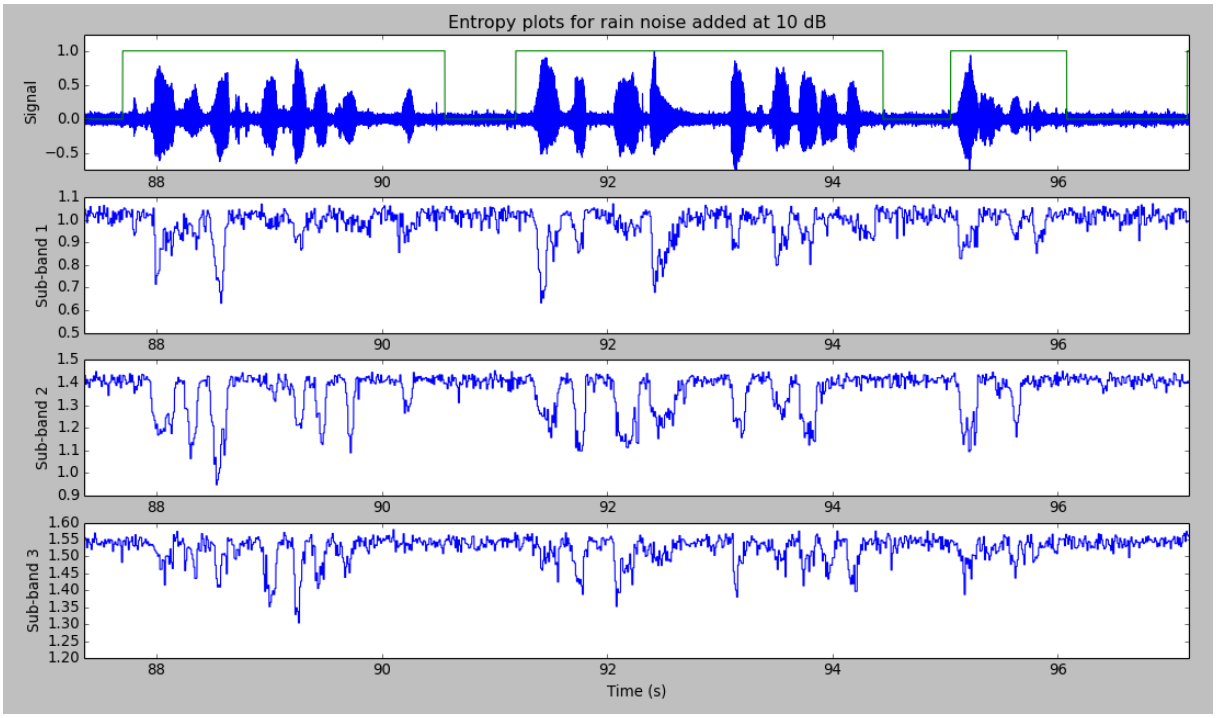

Figure 4.5: Example of entropy feature extracted for the case of rain noise. Green signal in the first plot indicates the groundtruth labels.

have ideally had on ZCR.

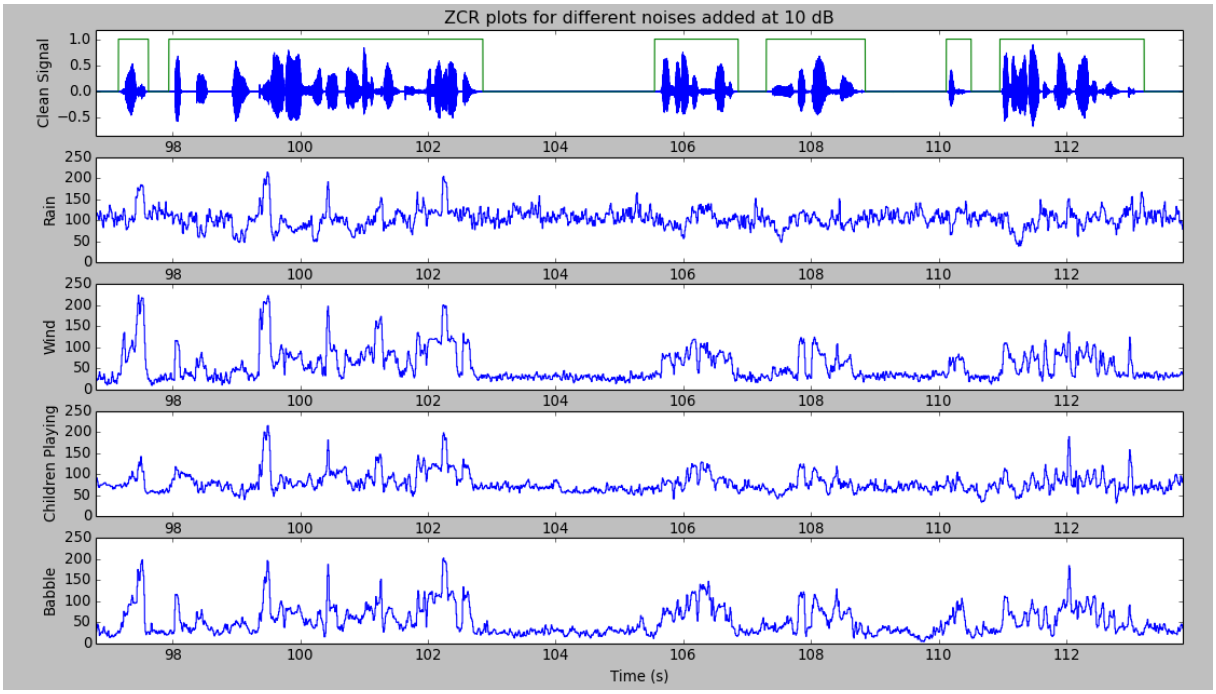

Figure 4.6: Zero-crossing rate extracted for different noise types. Green signal in the first plot indicates the groundtruth labels.

4.1.5 Modulation index

Amplitude modulation spectrogram (AMS), one of the popular features exploiting the modulation property of speech signals, faces an issue of being a high dimensional feature. The idea behind AMS is that the incoming signal is divided into n number sub-bands and then, in each of those n sub-bands, contribution (energy at the frequency bin) of y modulation frequencies is evaluated. This leads to a very high dimensional feature depending on n and y values. So, [3] investigates the effect of different speech frequency range (SFR) and modulation frequency range (MFR) in different noise scenarios. We adapt modulation index, their proposed feature, for our task by tuning the SFR and MFR parameters.

The input speech is passed through a band-pass filter of 300 Hz - 2000 Hz (SFR fixed for our task), and then the extracted envelope is passed through 16 different band-pass filters. The output of these latter filters belongs to has information about the modulation frequency. Details of the feature extraction are shown in figure 4.7. An MFR of 4 Hz to 16 Hz is chosen and framing is done in order to obtain a feature value per 10 milliseconds. And finally, the average of all the output signals obtained within the decided MFR is obtained as the modulation index. MFR and SFR are chosen by comparing the ROC plots generated by varying a constant threshold. As this feature exploits a very peculiar characteristic of speech, it is observed to be robust to babble noise and so it should also work with children playing noise, yet another challenging type.

From the plots in figure 4.8 it can be clearly seen that this feature is robust across different noise types; there is a very slight difference in the plot extracted from signals corrupted with different noise types. There are places where modulation index has failed to detect speech, mainly at the onsets and offsets, or in cases where speech segment is too small. This is where other features will aid its performance.

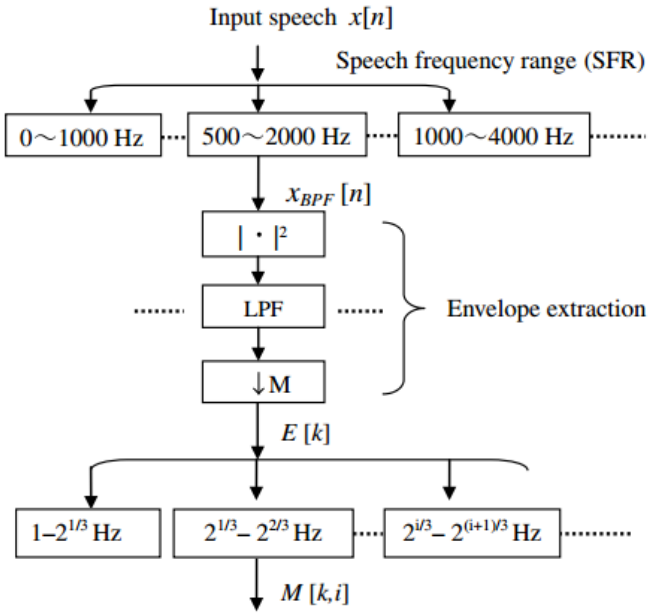

Figure 4.7: Extraction of modulation index feature [3]

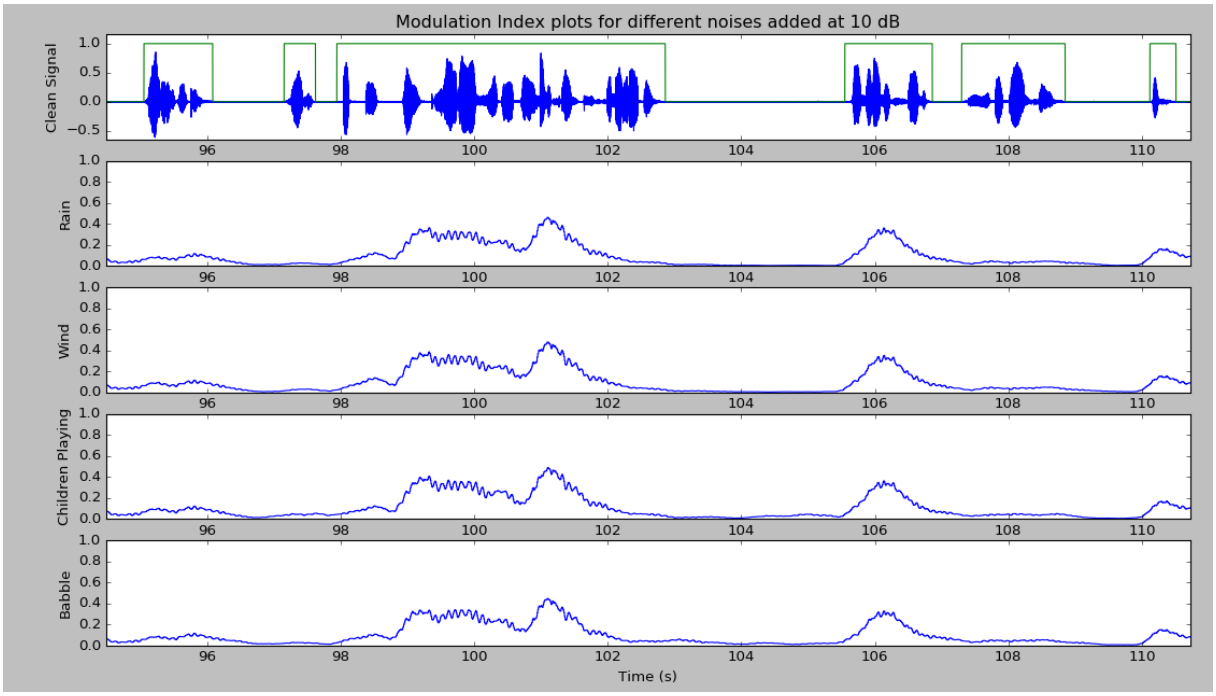

Figure 4.8: Modulation index extracted for different noise types. Green signal in the first plot indicates the groundtruth labels.

Table 4.1 summarizes the information about which feature targets what kinds of noises.

4.2 Classifier

Classification and regression tree (CART) is used as a classifier. It uses gini as the classification criteria, which is a measure of how often a randomly chosen element from the set would be

Feature	Rain	Wind	Babble	Children Playing
Energy-based (ALED)	✓			
Entropy sub-band 1,2	✓			✓
Entropy sub-band 3		✓		
Harmonicity-based (ZFF)	✓	✓		✓
Modulation Index	✓	✓	✓	✓
Zero-crossing rate		✓	✓	

Table 4.1: Summary of the features used and the respective target noise type

incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset [84]. While decoding, the threshold on the class prediction probability was decided using a validation set. The threshold corresponding to the hit rate of 0.85 is saved and used while decoding the test set. As the outcome of this VAD will decide what segments will be passed on to the ASR, a margin on the hit rate is set and not any other metric. The last thing we would want is losing the actual speech data, as it will lead to an increase in deletions in the ASR output. The final binary output is then saved for further processing.

4.3 Post-processing

4.3.1 Decision smoothing

The final decision output of these two methods is observed to have a large number of very small silence ($\leq$ 200 ms) and speech spurts ($\leq$ 100 ms), introduced because of the sensitivity of the features. Temporal smoothing (TS) was implemented in order to deal with such occurrences. We do an update of the raw decision output depending on the presence of the above-mentioned spurts. Taking all these points into consideration, a temporal smoothing algorithm (discussed in algorithm 2) is developed. This algorithm also preserves the release of a stop consonant by examining the duration of the short silence region just before that, which can otherwise get considered as a misdetected speech spurt. On applying TS on final output, huge improvements were observed in both frame-level as well as segment-level metrics. An example plot showing the outcome of decision smoothing has been shown in figure 4.9. In the algorithm 2,

1. 'segment' refers to an alternating 0,1 array, where all the frame-level contiguous zeros(ones) are mapped to a single zero(one),
2. 'noFrames' refers to an array having time duration of each segment in milliseconds,
3. 'update' function modifies the above two arrays after every iteration so that 'segment' retains its alternating 0,1 property.

4.4 Evaluation Criteria

In all we 6 evaluation metrics, all of which have been discussed in detail in section 3.3. For the results on task specific as well as synthesized data, we report the frame-level evaluation met-

Algorithm 2 Temporal Smoothing Function

```
1:  function TemporalSmoothing(rawDecision)
2:      j = 0
3:      [segment, duration] = convertFtoS(rawDecision)
4:      totalSegments = length(segment)
5:      while j < totalSegments do
6:          N = duration[j]
7:          if (segment[j] = 1 and N ≤ 100) then
8:              segment[j] ← 0
9:              update(segment, duration)
10:         if (segment[j] = 0 and N ≤ 200) then
11:             segment[j] ← 1
12:             update(segment, duration)
13:         totalSegments = length(segment)
14:         j ← j + 1
15:     decision = convertStoF(segment, duration)
16:     return decision
```

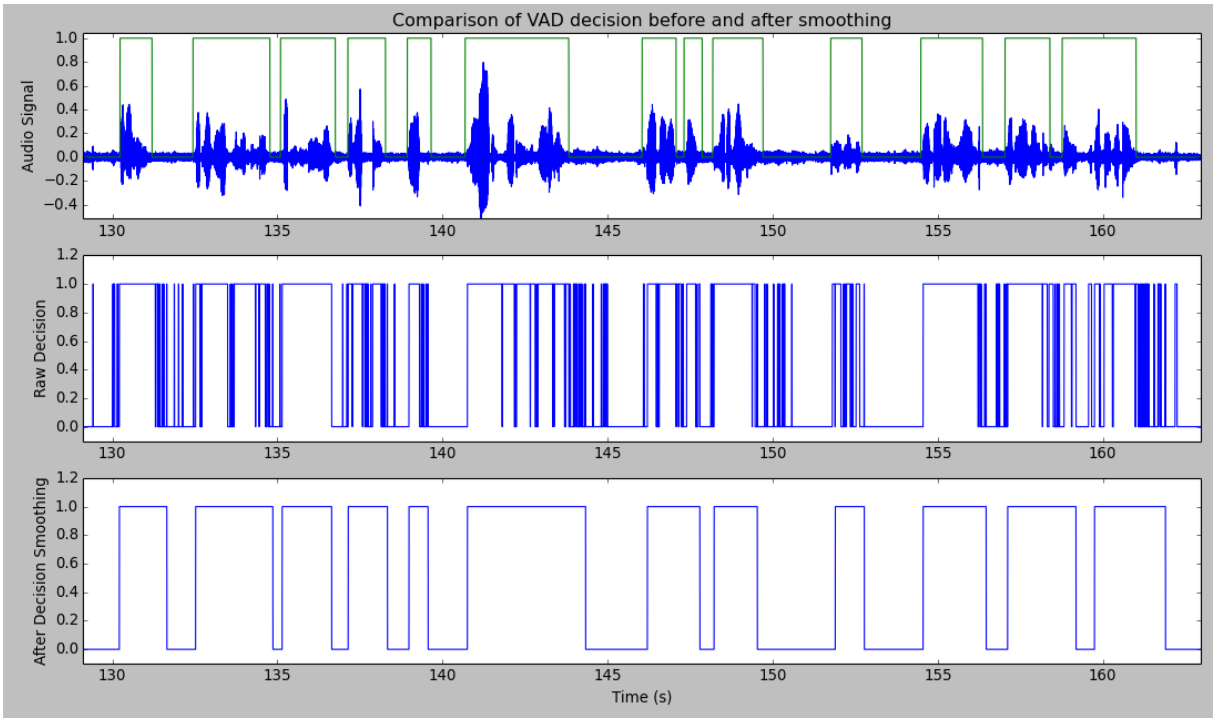

Figure 4.9: Improvements due to decision smoothing on raw VAD decision output. Green signal in the first plot indicates the groundtruth labels.

rics of FEC, MSC, OVER, NDS and overall accuracy, and a combined segment level accuracy metric. Segment level metric is reported in order to predict and compare the ASR performance on this data with respect to the benchmark algorithms, as we do not have this data annotated on word-level for getting ASR results. As we are using two benchmark algorithms, it is difficult to make any comparison using hit rate and false alarms rates. This is because for any such comparison we require at least one of the two quantities to be fixed to a certain value. As we

do not have any control over the benchmark algorithms, it is not possible to make the baselines match at a certain value. So the traditional evaluation metrics of HR and FA are not reported as it won't be possible to make any comparisons. Same will hold true for the first 4 metrics, direct comparison with two baselines might not be very informative. But it will be interesting to note what kind of errors contribute the most to the overall accuracy, and that is why those are reported. In our case, reporting accuracy makes sense because speech and non-speech are almost equally represented.

For the case of segmental metrics [78], we need to set the tolerance duration (refer section 3.3.2) for start boundary and end boundary accuracy parameters. This value decides how much delay in speech onset and how much early detection of speech offset is allowed for a non-zero accuracy of start and end boundary detection respectively. This value is set at a maximum of 500 milliseconds (50 frames) and a fifth of the length of the segment. The weighing function, in equations 3.12 and 3.14, is set to a unit step function.

For the data for ASR evaluation 2.4, we do not have the speech/non-speech ground truth labels. The finalized VAD algorithm is run on all the utterances. The VAD decision is smoothed according to the procedure mentioned in section 4.3.1. For comparison, SFF VAD is also run on all the utterances. Based on the VAD decision, all the detected speech segments are concatenated together while keeping a 50 milliseconds extra duration on both sides of each detected segment. The state-of-the-art ASR [21], proposed and tuned for our task-specific data, is run on these three sets audios (before and after modification), and phone error rate (PER) and other ASR-related metrics - insertion, deletion, substitution percentages, are reported.

Chapter 5

Results and Discussion

We have three kinds of dataset - synthesized noisy data, real-world noisy data, and dataset for ASR evaluation. All these datasets are processed through the proposed VAD and the outputs are used for further evaluation. The evaluation method is different for the three datasets as for the latter we do not have speech/non-speech ground-truths available but the dataset is appropriately labeled for ASR evaluation. The former two datasets are evaluated on 6 different evaluation metrics - overall frame-level accuracy, accuracy broken down into 4 other metrics depending on the position of the error within a segment and a combined segmental accuracy metric inspired from [78]. Reported are the values calculated across all the test files, 10 in the case of synthesized data, and 29 in the case of natural data. In the case of ASR results, the common metrics of phone error rate (PER), and individual percentage contribution of deletions, insertions, and substitutions have been reported. For natural data, HR and FA are reported too as the comparison is possible. While in the case of the synthesized data, no comparison was possible because whenever one of the metrics improved across different methods, the other would degrade. This is also explained in section 4.4. This phenomenon can be clearly seen from the bar graphs shown in figures 5.1 and 5.2. The trend across HR and FA is same in all the cases, except for babble noise where the difference is not that substantial either. This validates why using just HR and FA is not very informative. Segmental and frame-level accuracy, on the other hand, would give us an overall idea for comparison as they are just a single metrics in their respective domain.

5.1 VAD Results

The results on synthesized data have been reported in tables 5.1, 5.2, 5.3, and 5.4 for respective noise types and table 5.5 presents an average across all the noise-SNR combinations. The major observations along with plausible explanations have been described as follows -

1. AMR has the least segmental accuracy of all, even though the frame-level accuracy is comparable with other two methods. The reason behind this is a large number of segments present in the AMR's decision output, which is a consequence of poor decision smoothing involved (as it takes just 2 frames into account). An example plot is shown in figure 5.3. While doing a noise-wise comparison on the frame-level, we can observe that AMR performs well on rain, it being one of the stationary noises. It does not vary much temporally as well as spectrally throughout the audio. Next, it also performs well on wind noise as AMR considers energy in different subbands and wind, being a non-harmonic type of

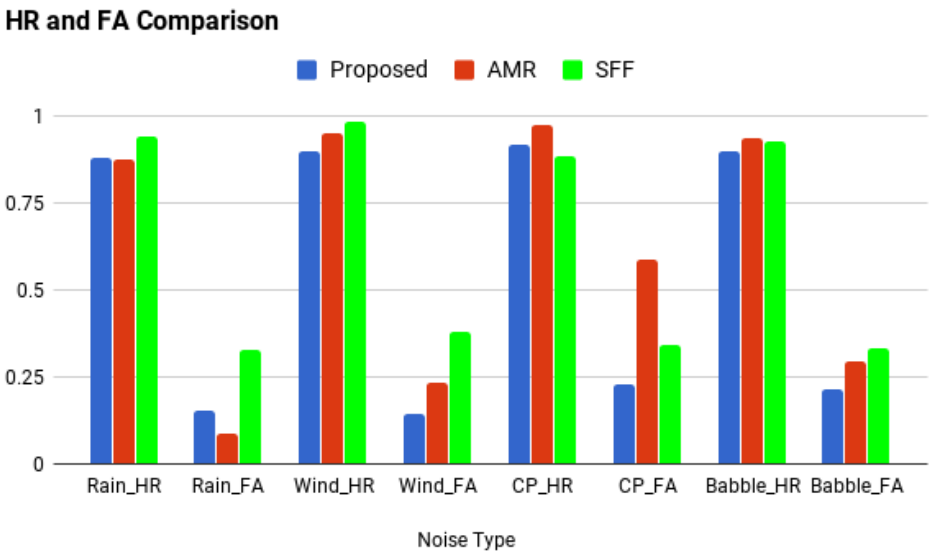

Figure 5.1: HR and FA compared across different methods. For the sake of simplicity results only on the 10dB case are shown.

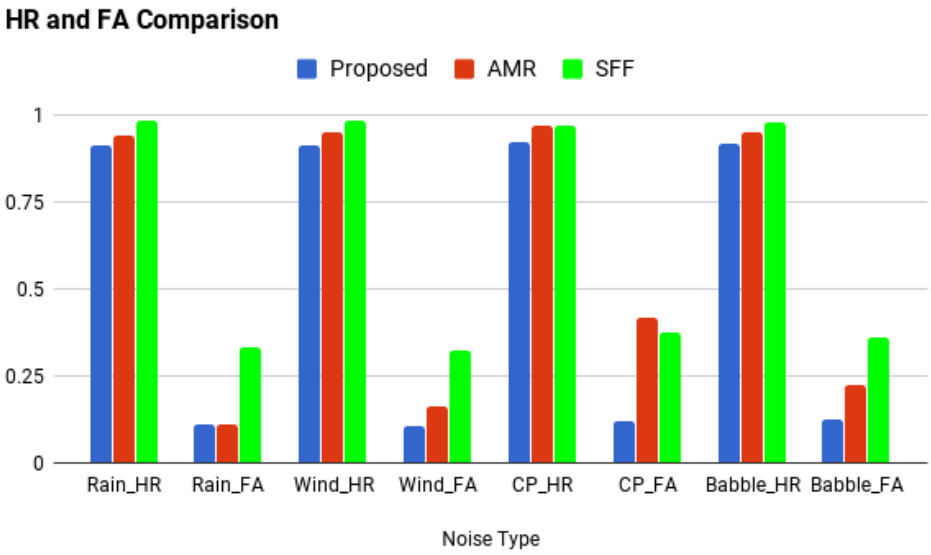

Figure 5.2: HR and FA compared across different methods. For the sake of simplicity results only on the 20dB case are shown.

sound, will not have a major contribution to any particular frequency band. Better performance on babble noise as compared to children playing could be accounted to the fact that babble is temporally (signal-energy) more stable as compared to the latter. Nonetheless, AMR's accuracy is lower than the proposed VAD for the all the noise types except rain as well as in terms of the overall average. The reason for the proposed VAD not performing as well as AMR on the rain could be because of the presence of features which degrade (ZCR, highest entropy subbands) in the presence of rain.

2. SFF and the proposed VAD have quite comparable segmental accuracies (within $\pm 5\%$); in

fact, it is consistently higher for SFF in the case of wind and rain noise in the synthesized data. Also, just for these two noise types, the majority of the SFF errors occur predominantly at the end of a speech segment (Over) and in the non-speech regions (NDS). From the definition of segmental metric we can infer that the errors at the beginning (FEC) and within (MSC) the speech segments will do more harm as they affect the start boundary accuracy and boundary precision respectively. So SFF's segment-level performance is better for these two noise types. The mentioned distribution of the four frame-level metrics implies a high HR and a high FA region of the ROC curve, which could also be seen from the bar charts in figures 5.1 and 5.2 that the SFF thresholding procedure has been designed to achieve a high HR but at a cost of high FA (which is constantly above 30%).

3. In terms of the frame-level accuracy, the proposed VAD outperforms SFF for every noise type. One reason for the poor frame-level accuracy could be because SFF exploits only half of the spectral information (till 4kHz) of that used by our proposed VAD (till 8kHz). SFF feature has a poor discrimination power for audios with babble noise, as can be inferred from the table for dynamic range provided by the authors [57]. Children playing being the more challenging noise, SFF's performance can be expected to be poor for these two cases (as is the observation). Also as these noise types are non-stationary, the mean of the lowest 20% feature values in the audio might be lower than most of the non-speech regions as well, thus giving rise to false alarms.

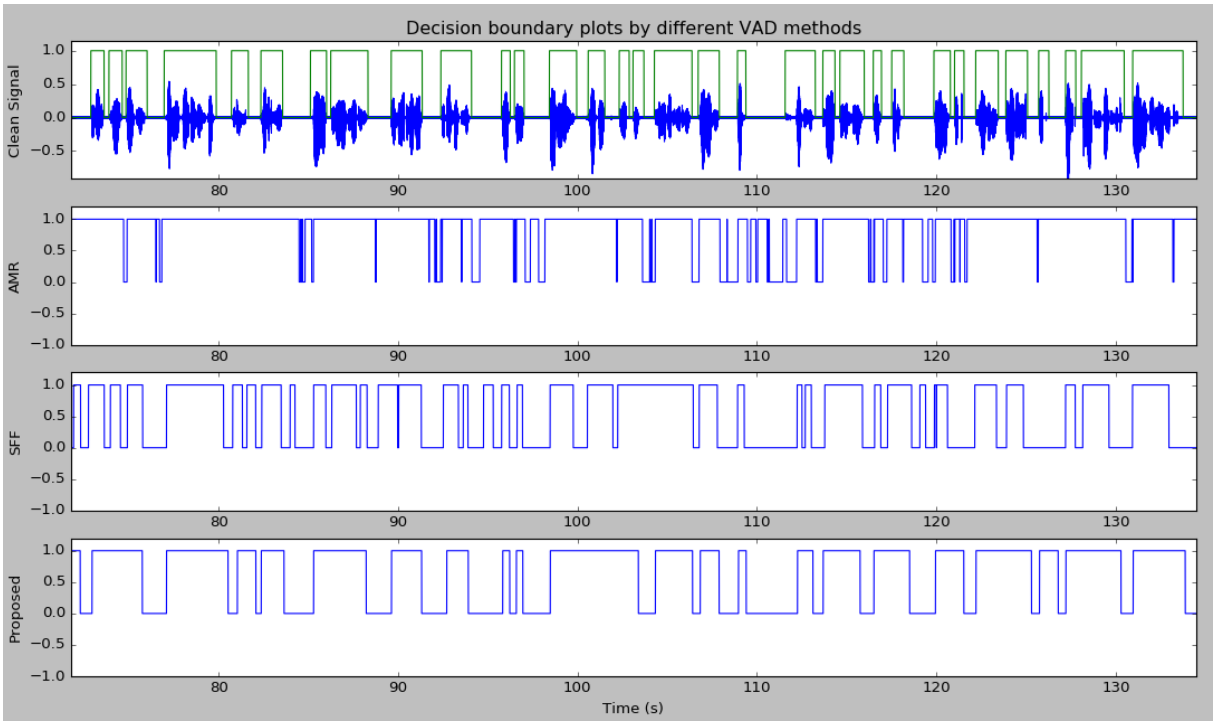

Figure 5.3: An example comparing decision across different VAD outputs for audio with children playing noise. Green signal in the first plot indicates the ground truth labels.

For natural data results reported in 5.6, it can be observed that HR and FA are better for the proposed VAD as compared to the SFF one. In the case of AMR, although the HR is higher by 3%, the FA rate is almost double that of the proposed VAD. Not only this, the proposed VAD

SNR	Method	Over	FEC	MSC	NDS	Accuracy	Segmental Metric
10	Proposed	0.065	0.031	0.034	0.004	0.865	0.87
	AMR	0.013	0.017	0.049	0.026	0.895	0.373
	SFF	0.1	0.006	0.025	0.052	0.818	0.925
20	Proposed	0.045	0.02	0.026	0.005	0.904	0.901
	AMR	0.021	0.011	0.021	0.027	0.918	0.453
	SFF	0.074	0.002	0.007	0.079	0.838	0.933

Table 5.1: Results for the case of rain noise

SNR	Method	Over	FEC	MSC	NDS	Accuracy	Segmental Metric
10	Proposed	0.06	0.026	0.029	0.005	0.88	0.891
	AMR	0.044	0.009	0.015	0.063	0.868	0.51
	SFF	0.082	0.002	0.007	0.093	0.817	0.94
20	Proposed	0.043	0.021	0.024	0.005	0.906	0.901
	AMR	0.029	0.01	0.015	0.045	0.901	0.542
	SFF	0.072	0.002	0.007	0.078	0.842	0.937

Table 5.2: Results for the case of wind noise

SNR	Method	Over	FEC	MSC	NDS	Accuracy	Segmental Metric
10	Proposed	0.092	0.023	0.022	0.013	0.85	0.917
	AMR	0.097	0.005	0.01	0.175	0.713	0.559
	SFF	0.056	0.014	0.047	0.101	0.782	0.779
20	Proposed	0.049	0.018	0.023	0.006	0.903	0.915
	AMR	0.055	0.006	0.009	0.136	0.792	0.474
	SFF	0.064	0.003	0.013	0.109	0.812	0.879

Table 5.3: Results for the case of children playing noise

SNR	Method	Over	FEC	MSC	NDS	Accuracy	Segmental Metric
10	Proposed	0.09	0.025	0.028	0.008	0.848	0.902
	AMR	0.039	0.011	0.023	0.097	0.829	0.475
	SFF	0.064	0.007	0.031	0.09	0.808	0.843
20	Proposed	0.052	0.019	0.024	0.006	0.898	0.907
	AMR	0.037	0.01	0.015	0.065	0.873	0.498
	SFF	0.069	0.002	0.009	0.097	0.823	0.902

Table 5.4: Results for the case of babble noise

Method	Frame-level accuracy	Segment-level accuracy
Proposed	0.88	0.90
AMR	0.84	0.48
SFF	0.81	0.89

Table 5.5: Average overall results for synthesized noisy data

Method	Over	FEC	MSC	NDS	Accuracy	HR	FA	Segmental Metric
Proposed	0.066	0.024	0.016	0.012	0.881	0.915	0.15	0.908
AMR	0.035	0.011	0.017	0.155	0.781	0.941	0.365	0.444
SFF	0.056	0.018	0.052	0.107	0.766	0.853	0.315	0.756

Table 5.6: Results for naturally noisy data

gives the better performance in terms of both segment-level as well as frame-level accuracy. The naturally noisy data has a good variety of noise types at different SNRs with an average of 17dB and the extreme being as low as 4.4dB. So on an average, the data is not as challenging as the synthesized one and so the VAD performs better than it performs on the synthesized data. We can see that most of the errors by proposed VAD are confined around the speech segments and the MSC and NDS values are low, which implies that the audio has not been over-segmented and thus the segmental metric has a better value. Higher FEC, on the other hand, is something which might pose a problem to the ASR performance. Children playing and babble noise are the most prominent ones in the natural scenario and both SFF and AMR have been observed to perform poorly on these two types, which reflects here as well. If the recording has different types of noises or is not evenly noisy, the threshold average will be set too low for some sections of noise, thus giving rise to false alarms; this holds for both, SFF as well as AMR.

5.1.1 Detecting extremely noisy environment

As mentioned before, we would also like to check if the recording environment is so noisy that we would have to eventually discard the recording. This will help save the reader's effort if we can prompt the user before he/she goes ahead and finishes the recording. In order to check whether the environment is too noisy, we can estimate the SNR from the VAD decisions and also use % speech information as obtained from the same. Given our application, we would expect the speech fraction to be around 30 to 70%. The credibility of the SNR estimate depends on how good the classification is and percentage of speech fraction in the audio is a very crude way to judge that. SNR estimate and the % speech duration was evaluated from the VAD decisions of the synthesized data. The values (SNR was either close to 10dB or 20dB and speech fraction was between 0.35 and 0.7) nearly agreed with the actual ones. Next, a few extremely noisy audios were considered. The estimated speech fraction was as low as 4.4% in one of the audios as no audible speech was recorded so almost everything got classified as noise. Another recording which had a very loud school noise throughout had an estimated speech fraction of 82% because the school noise also included babble and distinct background speaker in large parts which got misclassified as speech. When the speech estimate is within the acceptable range we can examine the SNR estimate to check if it is higher than a particular threshold level. For instance, in a recording which had a very high level of noise (the speech was audible) the speech fraction estimate was $\sim 50\%$, while the SNR estimate was just 5dB. Thus using the same logic for the first few seconds of the recording, we can judge if the environment noise levels are unacceptable and request the user to choose a quieter location. But this method still does not help the cases where the constant narrator audio in the background ruins the complete audio as the narrator audio, too, is present only in those half of the regions.

5.2 ASR Results

As we saw in the previous section, the segmental metric for SFF and the proposed VAD was comparable, significantly higher than that of AMR. So, this part of the analysis is done only on the data processed by the former 2 VAD techniques. Figures 5.4 and 5.5 show plots for example audios before and after pre-processing by both the VAD techniques. The regions which SFF VAD has failed to eliminate correspond to the babble noise in both the cases. In fact in figure 5.4 the extra signal is that of the narrator audio which the proposed VAD has successfully eliminated, this could be because of its lower amplitude. Thus we can see that the proposed VAD is successful in eliminating most of the background noise regions while ensuring that none of the speech regions are unnecessarily clipped off.

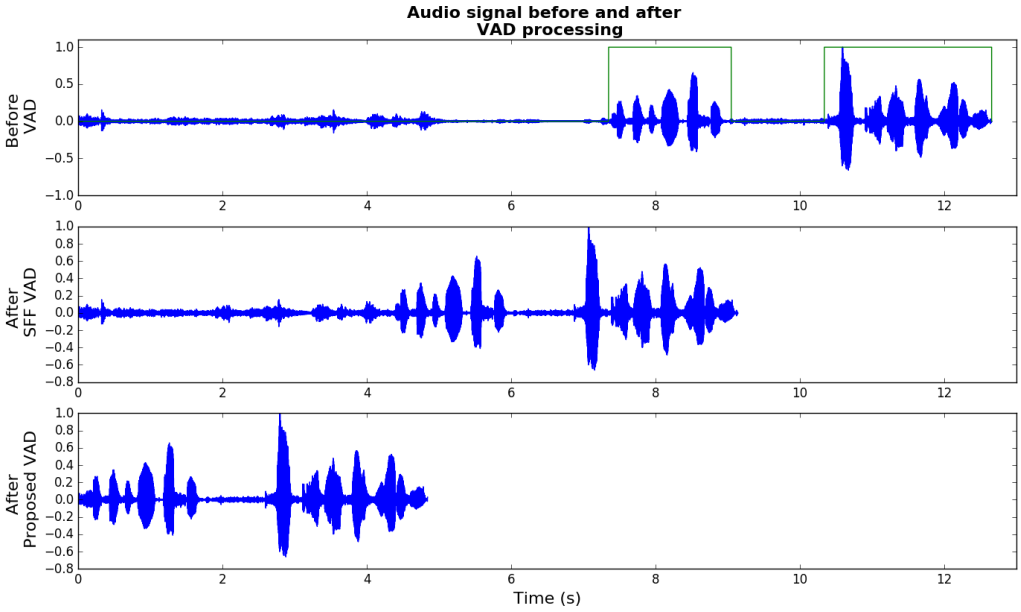

Figure 5.4: Plots for an utterance of noisy signal before and after VAD preprocessing. Green signal in the first plot indicates the speech signal region.

Table 5.7 summarizes the ASR results on the data with and without pre-processing through a VAD. The proposed VAD helps improve the performance in phone-error-rate (PER) by 6%. As would be expected of a VAD, the number of insertions have reduced, while the number of deletions and substitutions have almost remained the same. This implies that the VAD was successful in removing the unwanted non-speech regions while retaining the most of the speech part. It was observed that the VAD was able to remove stationary noise, school noise and soft babble noise in the background. The cases where our VAD miserably failed include the recordings with background talker, breath releases accentuated by a microphone, narrator audio in the background (a special case of background talker)and the audios with very low SNR.

SFF VAD, on the other hand, brings about no improvement. Although insertions and substitutions have reduced, deletions have gone up. This can only be because SFF VAD cut out the speech regions as well along with non-speech ones. The reason for this too lies in the thresholding criteria that SFF uses. As mentioned in section 2.4, each utterance is only about 5 seconds long on an average. So the assumption about 20% non-speech duration (which is used while

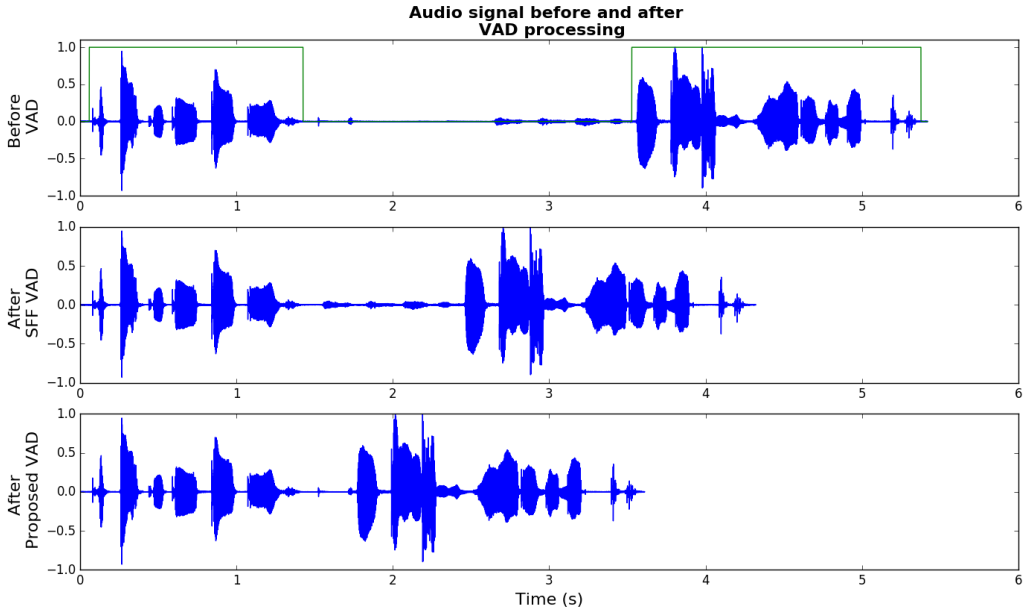

Figure 5.5: Plots for an utterance of noisy signal before and after VAD preprocessing. Green signal in the first plot indicates the speech signal region.

deciding the threshold) does not hold true in a significant number of cases. This is why in recordings which do not have that large % of non-speech regions, the threshold is set to a higher value, thus eliminating speech frames too. In fact, 8.6% of the 2951 audios have been completely wiped out by SFF VAD due to the same reason. All of these are very short duration recordings, about 1800 milliseconds each on an average. Shorter the recording lesser are the chances of having a significant non-speech duration. At this point, it would be interesting to note that the proposed VAD brings about a compression of 24%, while SFF VAD leads to a compression of 40%. Even though we do not know the optimum value of expected compression, this huge difference in compression percentages is yet another evidence that SFF might be cutting out speech-segments too along with the non-speech ones.

Method		**PER**	**Insertions(%)**	**Deletions(%)**	**Substitutions(%)**
without pre-processing		73.61	0.14	0.19	0.41
with pre-processing	SFF	74.37	0.07	0.36	0.31
	proposed VAD	67.72	0.09	0.2	0.39

Table 5.7: Comparison of ASR results on VAD pre-processed data

Thus the overall results are in accordance with the segment-level accuracy obtained for the natural data, as was also observed by Tong in [78].

Chapter 6

Conclusion and Future Work

We have adopted a classifier-based approach to the problem of voice activity detection in noisy speech. The motivation is to help the end-goal of automatic speech recognition in noisy conditions. We choose 8 features across different domains - energy, harmonicity, pitch, long-term, spectral. The noisy dataset was synthesized for the purpose and ease of generation of the labeled training data. Rain, wind, babble, children playing, were the four noise types considered on the basis of their relevance to the recording scenario and representation of a class of noise types. Decision smoothing on the obtained raw VAD output was observed to aid the accuracy to a great extent. Overall, a 4% and 7% improvement in frame-level accuracy in VAD performance over benchmark algorithms is observed (table 5.5). Segmental accuracy almost matches the better performing benchmark. Segmental accuracy measure has been used for comparison as it not only provides a single evaluation measure which leads to simplicity in comparison but also is known to be correlated with the ASR accuracy. The proposed VAD technique is observed to improve the ASR performance by 6% in terms of phone error rate; having said that, we cannot comment on how good this improvement is.

We still have not used any cepstral-domain feature, which could lead to an improvement as it captures another dominant characteristic of speech signals. MFCC, which is motivated by the human hearing mechanism and focuses on the formant structure of the speech signal, has been previously and successfully used as a feature in the classifier-based approach towards VAD [54, 56]. The proposed VAD uses entropy which does partially depend on the formant structure, but it does not contribute much to the final classifier decision (as can be seen from the relative feature importances). So a feature which more directly depends on this unique characteristic of speech might be of greater help. Another feature, cepstral peak, derived from the higher order bins of cepstrum, has been shown to be a superior to the traditionally used harmonicity-based features like zero-crossing rate, auto-correlation function [10]. The classifier, too, can be improved to introduce a non-linearity in the classification boundaries. At present, the system is a bit deviant from that proposed in the block diagram 2.4, as two VADs are in place. The first one is an energy-based VAD applied before uploading on the interface and then a classifier-based VAD is applied before passing on the audios to an ASR engine. Because of this, our comparison with the benchmark is not full proof yet. If SFF had been run right before uploading the audios to the rating interface, it wouldn't have faced the problem of having less % of non-speech segments than the underlying assumption. We need to incorporate that in order to do an accurate end-to-end comparison. Also, in order to evaluate whether the correlation between segmental

accuracy and ASR accuracy does exist, we need a common dataset labeled for VAD as well as ASR. Availability of such dataset will also help us quantify how good the VAD algorithm is, how much % of the total non-speech frames has the algorithm successfully eliminated.

References

[1] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[2] ETSI, "Voice activity detector for adaptive multi-rate speech traffic channels," *ETSI EN 301 708 V7.1.0*, 1999.

[3] K. Pek, T. Arai, and N. Kanedera, "Voice activity detection in noise using modulation spectrum of speech: Investigation of speech frequency and modulation frequency ranges," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 33–44, 2012.

[4] "ASER: The Annual Status of Education Report (rural)," http://www.asercentre.org//p/289.html, ASER Centre.

[5] "ASER: The Annual Status of Education Report (rural)," http://img.asercentre.org/docs/Publications/ASER%20Reports/ASER%202016/aser_2016.pdf, ASER Centre, 2016.

[6] "LETS : Learn English Through Stories (2016)," http://www.tatacentre.iitb.ac.in/15mobitech.php, Tata Centre for Design and Technology at IIT Bombay.

[7] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[8] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[9] J. Kola, C. Espy-Wilson, and T. Pruthi, "Voice activity detection," Maryland Engineering Research Internship Best Project Award, Tech. Rep., 2011.

[10] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Features for voice activity detection: a comparative analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, p. 91, 2015.

[11] N. Kitaoka, K. Yamamoto, T. Kusamizu, S. Nakagawa, T. Yamada, S. Tsuge, C. Miyajima, T. Nishiura, M. Nakayama, Y. Denda *et al.*, "Development of vad evaluation framework censrec-1-c and investigation of relationship between vad and speech recognition performance," in *Proc. of IEEE Workshop on Automatic Speech Recognition & Understanding*, 2007.

[12] "Sensibol reading tutor app (2016)," http://sensibol.com/readingtutor.html, SensiBol Audio Technologies Pvt. Ltd.

[13] "Bookbox: A book for every child in her language," www.bookbox.com, bookBox.

[14] P. Rao, P. Swarup, A. Pasad, H. Tulsiani, and G. G. Das, "Automatic assessment of reading with speech recognition technology," in *Proc. of International Conference on Computers in Education*, 2016.

[15] "University of Toronto," https://tspace.library.utoronto.ca/handle/1807/66306, The Natural Sound Library, 2014.

[16] D. Sahi, *Essential Indian Sound Effects*. Parvati Pictures Pvt. Ltd., 1999.

[17] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[18] "Youtube," https://www.youtube.com, youTube: Noise Audios.

[19] "Nature Sounds for Me," http://naturesoundsfor.me/, nature Sounds Database.

[20] "RWTH Aachen University," http://www.iks.rwth-aachen.de/en/research/tools-downloads/wind-noise-database/, wind Noise Database.

[21] K. Sabu, P. Swarup, H. Tulsiani, and P. Rao, "Automatic assessment of children's l2 reading for accuracy and fluency," in *Proc. of Speech and Language Technology in Education*, 2017.

[22] E. Chuangsuwanich and J. R. Glass, "Robust voice activity detector for real world applications using harmonicity and modulation frequency," in *Proc. of INTERSPEECH*, 2011.

[23] Q. Li, J. Zheng, Q. Zhou, and C.-H. Lee, "Robust, real-time endpoint detector with energy normalization for asr in adverse environments," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[24] P. C. Khoa, "Noise robust voice activity detection," Ph.D. dissertation, Nanyang Technological University, 2012.

[25] "A silence compression scheme for g.729 optimized for terminals conforming to itu-t v.70," *ITU-T Recommendation G.729, Annex B*, 1996.

[26] L.-s. Huang and C.-h. Yang, "A novel approach to robust speech endpoint detection in car environments," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, 2000.

[27] K. Sakhnov, E. Verteletskaya, and B. Simak, "Dynamical energy-based speech/silence detector for speech enhancement applications," in *Proc. of World Congress on Engineering*, 2009.

[28] K.-H. Woo, T.-Y. Yang, K.-J. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[29] N. Cho and E.-K. Kim, "Enhanced voice activity detection using acoustic event detection and classification," *IEEE Trans. on Consumer Electronics*, vol. 57, no. 1, 2011.

[30] R. V. Prasad, A. Sangwan, H. Jamadagni, M. Chiranth, R. Sah, and V. Gaurav, "Comparison of voice activity detection algorithms for voip," in *Proc. of International Symposium on Computers and Communications*, 2002.

[31] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[32] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3, pp. 271–287, 2004.

[33] P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," 2001.

[34] J. Ramirez, J. C. Segura, C. Benitez, A. de La Torre, and A. Rubio, "Voice activity detection with noise reduction and long-term spectral divergence estimation," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[35] T. Fukuda, O. Ichikawa, and M. Nishimura, "Long-term spectro-temporal and static harmonic features for voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 834–844, 2010.

[36] J.-l. Shen, J.-w. Hung, and L.-s. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments." in *Proc. of International Conference on Spoken Language Processing*, 1998.

[37] N. Madhu, "Note on measures for spectral flatness," *Electronics letters*, vol. 45, no. 23, pp. 1195–1196, 2009.

[38] S. Basu, "A linked-hmm model for robust voicing and speech detection," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[39] D. J. Nelson and J. Pencak, "Pitch-based methods for speech detection and automatic frequency recovery," in *Proc. of International Symposium on Optical Science, Engineering, and Instrumentation*, 1995.

[40] G. Hu and D. Wang, "Segregation of unvoiced speech from nonspeech interference," *Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. 1306–1319, 2008.

[41] A. Pasad, K. Sabu, and P. Rao, "Voice activity detection for children's read speech recognition in noisy conditions," in *Proc. of National Conference on Communications*, 2017.

[42] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," 2005.

[43] K. Deepak, B. D. Sarma, and S. M. Prasanna, "Foreground speech segmentation using zero frequency filtered signal," in *Proc. of Conference of the International Speech Communication Association*, 2012.

[44] K. R. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions,"

in *Proc. of International Symposium on Intelligent Signal Processing and Communication System*, 2000.

[45] A. Prathosh, T. Ananthapadmanabha, and A. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2471–2480, 2013.

[46] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[47] A. Tsiartas, T. Chaspari, N. Katsamanis, P. K. Ghosh, M. Li, M. Van Segbroeck, A. Potamianos, and S. Narayanan, "Multi-band long-term signal variability features for robust voice activity detection," in *Proc. of INTERSPEECH*, 2013.

[48] A. M. Liberman, *Speech: A special code*. MIT press, 1996.

[49] Y. Ma and A. Nishihara, "Efficient voice activity detection algorithm using long-term spectral flatness measure," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, p. 87, 2013.

[50] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.

[51] J.-H. Bach, B. Kollmeier, and J. Anemüller, "Modulation-based detection of speech in real background noise: Generalization to novel background classes," in *Proc. of International Conference on Acoustics Speech and Signal Processing*, 2010.

[52] H. You and A. Alwan, "Temporal modulation processing of speech signals for noise robust asr," in *Proc. of INTERSPEECH*, 2009.

[53] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[54] T. Kinnunen, K.-A. Lee, and H. Li, "Dimension reduction of the modulation spectrogram for speaker verification," in *Proc. of the Speaker and Language Recognition Workshop*, 2008.

[55] K. Ishizuka and T. Nakatani, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio." in *Proc. of INTERSPEECH*, 2006.

[56] A. Martin, D. Charlet, and L. Mauuary, "Robust speech/non-speech detection using lda applied to mfcc," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2001.

[57] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 23, no. 4, pp. 705–717, 2015.

[58] W. Ong and A. Tan, "Robust voice activity detection using gammatone filtering and entropy," in *Proc. of International Conference on Robotics, Automation and Sciences*, 2016.

[59] X.-L. Zhang and D. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection." in *Proc. of INTERSPEECH*, 2014.

[60] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound.* MIT press, 1994.

[61] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 1998.

[62] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 412–424, 2006.

[63] J.-H. Chang and N. S. Kim, "Voice activity detection based on complex laplacian model," *Electronics Letters*, vol. 39, no. 7, pp. 632–634, 2003.

[64] J.-H. Chang, J. Shin, and N. Kim, "Voice activity detector employing generalised gaussian distribution," *Electronics Letters*, vol. 40, no. 24, pp. 1561–1563, 2004.

[65] J. W. Shin, J.-H. Chang, and N. S. Kim, "Statistical modeling of speech signals based on generalized gamma distribution," *IEEE Signal Processing Letters*, vol. 12, no. 3, pp. 258–261, 2005.

[66] J. Tchorz and B. Kollmeier, "Snr estimation based on amplitude modulation analysis with applications to noise suppression," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 3, pp. 184–192, 2003.

[67] J. Wu and X.-L. Zhang, "Efficient multiple kernel support vector machine based voice activity detection," *IEEE Signal Processing Letters*, vol. 18, no. 8, pp. 466–469, 2011.

[68] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice activity detection using mfcc features and support vector machine," in *Proc. of International Conference on Speech and Computer*, 2007.

[69] J. Wu and X.-L. Zhang, "Maximum margin clustering based statistical vad with multiple observation compound feature," *IEEE Signal Processing Letters*, vol. 18, no. 5, pp. 283–286, 2011.

[70] M. Van Segbroeck, A. Tsiartas, and S. Narayanan, "A robust frontend for vad: exploiting contextual, discriminative and spectral cues of human voice." in *Proc. of INTERSPEECH*, 2013, pp. 704–708.

[71] T. V. Pham, C. T. Tang, and M. Stadtschnitzer, "Using artificial neural network for robust voice activity detection under adverse conditions," in *Proc. of International Conference on Computing and Communication Technologies*, 2009.

[72] O.-W. Kwon and T.-W. Lee, "Optimizing speech/non-speech classifier design using adaboost," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[73] T. Usukura and W. Mitsuhashi, "Voice activity detection using adaboost with multi-frame information," in *Proc. of International Conference on Signal Processing and Communication Systems*, 2008.

[74] M. Farsinejad and M. Analoui, "A new robust voice activity detection method based on genetic algorithm," in *Proc. of Telecommunication Networks and Applications Conference*, 2008.

[75] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.

[76] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[77] D. Freeman, G. Cosier, C. Southcott, and I. Boyd, "The voice activity detector for the pan-european digital cellular mobile telephone service," in *Proc. International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1989.

[78] S. Tong, N. Chen, Y. Qian, and K. Yu, "Evaluating vad for automatic speech recognition," in *Proc. of International Conference on Signal Processing*, 2014.

[79] P. C. Khoa and C. E. Siong, "Spectral local harmonicity feature for voice activity detection," in *Proc. of International Conference on Audio, Language and Image Processing*, 2012.

[80] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano, "Performance evaluation and comparison of g. 729/amr/fuzzy voice activity detectors," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 85–88, 2002.

[81] "Ansi-c code for the floating-point adaptive multi-rate (amr) speech codec," https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1400, version 13.0.0 (2015).

[82] J. C. Catford, *A practical introduction to phonetics*. Clarendon Press Oxford, 1988.

[83] Y. Faycal and M. Bensebti, "Comparative performance study of several features for voiced/non-voiced classification," *International Arab Journal of Information Technology*, vol. 11, no. 3, pp. 293–299, 2014.

[84] "Gini impurity criteria," https://en.wikipedia.org/wiki/Decision_tree_learning#Gini_impurity.