Predicting lexical skills from oral reading with acoustic measures

Dual Degree Project

Stage 2 Report

Submitted in partial fulfillment of the requirements for

Dual Degree (B.Tech + M.Tech)

by

Charvi Vitthal

(Roll No. 140070022)

Under the guidance of

Prof. Preeti Rao

Department of Electrical Engineering

Indian Institute of Technology Bombay

June 2019

Approval Sheet

This is to certify that the dissertation titled on **Predicting lexical skills from oral reading with acoustic measures** by **Charvi Vitthal (140070022)** is approved for the degree of **Dual Degree (B.Tech. + M.Tech)** in Electrical Engineering with a specialization in **Communication & Signal Processing**.

Examiner1 : Signature: Ashish Panda.

Examiner2 : Signature: V. Ri B...

Chairperson: Signature: V. Ri B...

Supervisor : Signature: Preeti Rao

CHARVI VITTHAL
20 JUNE 2019

Declaration of Academic Ethics

I declare that this written submission represents my ideas in my own words. 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.

Charvi Vitthal
(Roll: 140070022)
Department of Electrical Engineering
IIT Bombay

Date: June 4, 2019

Acknowledgement

I express my deepest gratitude towards my guide, Prof. Preeti Rao, for providing me the constant support and guidance throughout the project. I am thankful to Kamini Sabu and Shreeharsha B.S. for their collaboration in the project. I am also thankful to my friends, especially Nischal Agrawal, for the insightful discussions and other help.

Charvi Vitthal
Electrical Engineering
IIT Bombay

Abstract

Literacy assessment is an essential activity for education administrators across the globe. Typically achieved in the school setting by testing a child's oral reading, it is intensive in human resources. While automatic speech recognition (ASR) technology is a potential solution to the problem, it tends to be computationally expensive and also unreliable for any conditions that deviate from those of the training data set utilized in the ASR acoustic and language model building. A fluency measure that can be computed directly on the acoustic signal can be beneficial. In this work, we investigate such a measure and evaluate it in terms of the accuracy of prediction of reading skill as quantified by read text that is manually labeled at the word level. We first automatically cluster readers based on the lexical characteristics typical of different skill levels. Unsupervised clustering reveals the following dominant classes: good readers (mostly correct words, some disfluencies, i.e., self-corrections), poor readers (missed and/or unintelligible speech). We next propose a set of acoustic signal features to predict the word-decoding skill categories of interest from the recorded speech. Pause statistics, syllable rate, and spectral and intensity dynamics are found to be reliable indicators of specific types of oral reading deficits, providing useful feedback by discriminating the different characteristics of beginning readers. Note that these features do not require ASR or word boundaries to be known.

Contents

List of Figures

Chapter 1

Introduction

Learning a language requires good feedback in both quality and quantity. Reading aloud has traditionally been an essential instructional component in school curricula. Further, oral reading can serve to evaluate, both a child's word decoding ability and text comprehension [2]. The absence, or minimal occurrence, of word-level miscues such as deletions, substitutions, and disfluencies reveals good word decoding skills. On the other hand, comprehension is indicated by prosodic fluency, which a child typically acquires after word decoding becomes easy enough to free up the necessary cognitive resources [3, 4, 2]. Assessment based on oral reading involves having an expert (such as a language teacher) listen to the child reading a chosen text for attributes such as speech rate, correctly uttered words, phrasing, and expressiveness. It is thus intensive in human resources. As per ASER 2018, an annual survey by the education non-profit organization Pratham, [5], only half (50.3 %) of the students in fifth grade (class V) in rural India can read texts meant for the second grade (class II). One main reason for this is the dismal student to teacher ratio due to which students are not given the required level of attention.

There have been attempts to use ASR to evaluate lexical miscues followed by automatic analyses of the word-level segmentations for prosody evaluation [6, 7, 8]. In the reading context, ASR benefits from language models tuned to the intended text. On the other hand, due to the sensitivity to acoustic model training data mismatch, ASR is successful only when the speaker and environment variability is controlled. In the school scenario, diversity in skill levels and accents and, possibly also, background noise presence, affect the performance of both the language and acoustic models (LM and AM) making ASR unreliable, especially for the poor word decoders.

In this work, we investigate the possibility of using acoustic signal analyses to achieve a coarse categorization of word-decoding (i.e., lexical) abilities. The categories are obtained by clustering frequencies of the specific types of lexical miscues in a corpus of manual transcriptions of a large number of instances of distinct stories read by children across reading skill levels. We propose to detect instances of poor word-decoding using the appropriate cues obtained purely by acoustic signal analyses. It is expected that ASR based evaluation would be useful as a second stage only for the good instances in terms of expected near-adherence to the intended text. We thus expect to also benefit from the computational savings of acoustic analyses over ASR based analyses for a significant component of the test data.

Traditionally, word decoding skill is defined by the WCPM (words correct per minute) as measured by listening to the read-out text. More recently, the manual scoring of fluency has also been considered based on perceived phrasing and expressiveness [9]. Most research groups working in automatic reading feedback and assessment have focused on improving the performance of an ASR module for reading miscue detection and speech rate measurement [10, 11, 12, 13, 14]. Research has also addressed reading skill prediction using lexical [10] or prosodic [15] features or a combination [16] of both. Bolanos et al. [16] consider a 2-stage evaluation, where the first stage separates the poor readers from good readers, while the second stage discriminates the prosodically good from poor. The features used in both stages are drawn from the same set of acoustic and lexical features. In the present work, on the other hand, we propose an initial screening stage involving lexical skill prediction with acoustic analysis alone.

Previous work has exploited fluency metrics such as speaking rate and pause lengths and frequencies to predict non-native adult speaker proficiencies in communication settings [17]. Fontan et al. [18] used low-level signal features to estimate speech rate and its regularity in order to predict human ratings of fluency for Japanese learners of French. While there are other examples of the correlation of computed acoustic signal features with human expert rated fluency, none of these works attempt to predict specific word decoding attributes from the prosodic analyses.

In the present work, we categorize a typical dataset of children's read speech recordings into classes discovered by the unsupervised clustering of the observed lexical miscues in the manually generated transcriptions. We next investigate acoustic measures to predict the so identified broad categories with a view to develop an automatic system that provides useful descriptions of overall lexical skill. Experimental evaluation of the performance of the system is followed by a critical discussion of the results.

Chapter 2

Dataset

The data used in our study consists of recordings of short stories, by students in the age group 9-13 years via an app designed in-house. The students are learning English as an additional language and studying in 6th-8th standards in schools in rural areas. As mentioned above, our evaluation system could be particularly useful for children in rural areas where there is a lack of skilled English teachers and other learning and evaluation facilities like the internet. For the same reason, we expect a more extensive range of reading proficiency among students in rural areas, whereas those in the urban areas are expected to have mostly good proficiency level. The story content is tailored to be easily read by an average 6-8-year-old, but we see a rather wide range of reading proficiency in our data due to the child's acute lack of exposure to the language.

These stories are present in a video karaoke form, provided by Bookbox [19], and were set up on a specially created android app for recording by Sensibol [20]. This app presents the story in video karaoke mode with roughly one sentence per video screen. The student also has the option to shadow the reference speaker, i.e., attempt to mimic the speech of the reference audio. This feature can give rise to the interference of the reference audio with the student's speech in some of the recordings. A snapshot of the app is present in Figure 2.1. The words are highlighted in the sequence corresponding to an average speaking rate, i.e., that of the reference audio. At the end of each sentence, the video screen switches to that corresponding to the next sentence. All the recordings are made at 16 kHz sampling frequency with a headset mic to minimize background noise and are stored with metadata comprising the child's credentials, story name, and date of recording. This recording is then uploaded on a web-based interface, created by Sensibol, after appropriate sentence-level segmentation. The information of time-stamps of where

Figure 2.1: Snapshot of the mobile app with the video karaoke form of recording.

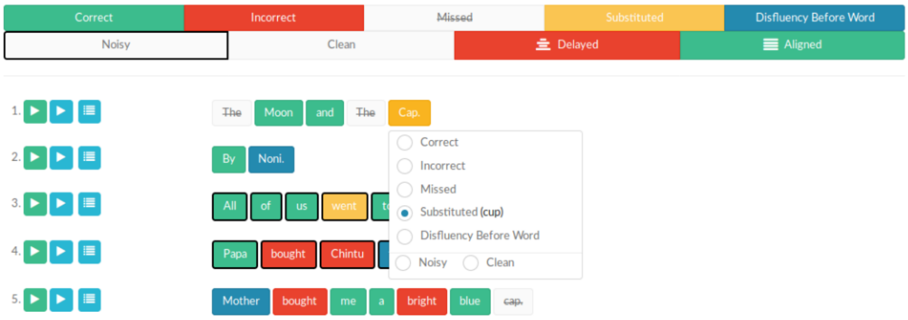


Figure 2.2: Manual transcription UI with colour-coded lexical miscues marked.

the video screen changes and a Voice Activity Detector (VAD) are used to segment the sentences.

These segmented sentences of the audio, aligned with the intended story text are provided on the web-interface to facilitate manual transcription at the word level. Figure2.2 shows the web-based interface for manual transcriptions. A human transcriber carries out word-level transcriptions using these segmented sentence-level recordings. Here, each canonical text word is compared to the same word in the reference audio and marked in different dimensions using color codes, as indicated in table2.1. Apart from the indicators mentioned, the transcriber can also make an additional comment, for example, on the type and level of noise. A stage of validation of these transcriptions follows next.

The substitutions mentioned in table 2.1 can be by one or more words. The substituted word(s) are keyed in by the transcriber for future use in Automatic Speech Recognition (ASR) Acoustic Model or Language Model training. Gibberish speech, marked red in the manual transcriptions, is observed to occur typically as a sequence of several syllables with no clear one-to-one correspondence to the expected words. This phenomenon is expected to be difficult to discriminate from correct speech without ASR.

Table 2.1: Indicators used for manual transcription

Indicator on the Interface	Description
Green	Correct - the word is pronounced correctly
Blue	Disfluency - the word is uttered in incomplete form or/and immediately corrected
Yellow	Substituted - the word is perceived as a different word(s)
Red	Incorrect - unintelligible speech or gibberish
Word is struck off	Missed - the word is skipped, i.e. not uttered at all

As the recordings were taken in the school environment, they contain a lot of background noise typically found in a school environment. The noise level ranges from fairly clean to very noisy. The different noise types are summarized in table 2.2.

Table 2.2: Different noise types seen in our data

Noise type	Descriptive characteristics
White noise, Wind, Rain	variable intensity, white noise like spectra
School Noise (children playing in background, bell noise, etc.)	variable spectra
Background Talker, Babble	non-stationary, speech-like

For our experiments, we select recordings belonging to 6 stories out of the total 18 available. These stories were selected to be of moderate difficulty level, to contain a significant number of corresponding audio recordings and to represent variability in the vocabulary. If the vocabulary is very repetitive, the student might learn to speak only the recurring phrase correctly and give a false indication of higher reading proficiency by speaking a significant fraction of the story correctly. For example, the phrase "Who does it all? Vayu, the wind." is repeated after every one-two sentences in the story "Vayu, the Wind".

The data used in our experiments includes 1090 recordings of 6 distinct stories, recorded by 212 distinct students. Each story is between 10-40 sentences long. The recording duration ranges from approximately 1 min to 6 min long. Detailed information about this subset of data used in our experiments is mentioned in the Table 2.3.

In some recordings, after the student has finished reading the text, he/she can be heard talking to the supervisor or to his/her friends in their L1 language, which is usually Hindi, Marathi or Telugu in our case. To ignore this unwanted speech, we use the recording only until the last video frame is present. As the recordings are supervised, we assume that

Table 2.3: Data distribution when clustering is performed at the sentence group level

Story	# recordings	# distinct students	# sentences per story	story video duration (min)
The moon and the cap	322	115	21	1.5
The lion and the fox	153	96	22	3.9
The flying elephant	80	52	41	6
The musical donkey	93	59	29	2.6
Bunty and bubbly	387	109	16	2.1
The wind and the sun	54	36	43	5.9

such unwanted speech is not present during the story video is playing.

Chapter 3

Lexical Clustering

We apply unsupervised clustering to discover possible underlying groupings of the students. For this, we use K-means clustering algorithm implemented in the sklearn library in python [21], and evaluate the system for different choices of K, i.e., the number of classes. Section 3.1 describes the different input features used and section 3.2 gives a brief recap of the K-means clustering algorithm. The quality of fit is evaluated using the silhouette score, which is explained in section 3.3. Section 3.4 contains discussion on the obtained results.

3.1 Input Features

We use the manual transcriptions, described in table 2.1, for categorizing the students based on their English reading proficiency. The proportion of each audio-indicator present in table 2.1, is computed and used as a feature for clustering the students. The audio recording corresponding to the time interval for which this feature is computed is called an "utterance and details about the computation follow below. An "utterance" can be defined in different ways. For example, a single sentence or a group of sentences or the whole story can be used as an "utterance". More discussion on the best choice of utterance is present in section 3.4.

The fraction of the different types of words transcribed on the Sensibol interface, table 2.1, with respect to the total words in the utterance is used to categorize the student. These, along with the terminology used henceforth, have been explained below :

- C = fraction of correct words in the utterance
  $= \dfrac{\text{total \# correct words in the utterance}}{\text{total \# words in the utterance}}$

- M = fraction of missed words in the utterance
  $= \dfrac{\text{total \# missed words in the utterance}}{\text{total \# words in the utterance}}$
- I = fraction of incorrect (gibberish) words in the utterance
  $= \dfrac{\text{total \# incorrect (gibberish) words in the utterance}}{\text{total \# words in the utterance}}$
- D = fraction of disfluent words in the utterance
  $= \dfrac{\text{total \# disfluent words in the utterance}}{\text{total \# words in the utterance}}$
- S_k = fraction of words substituted by k words in the utterance
  $= \dfrac{\text{total \# words substituted by k words in the utterance}}{\text{total \# words in the utterance}}$

By the transcriptions on the Sensibol interface, we have information about S_1, S_2, S_3 and $S_{>3}$.

Considering all the fractions mentioned above, we obtain an 8-dimensional feature vector. Along with this, we also examine some of the feature vectors obtained by meaningful combinations of above fractions. Although we have information about S_1, S_2, S_3 and $S_{>3}$ individually, their occurrence is very low and also, considering these 4 kinds of substitutions separately does not contribute to our task of categorizing the students, any more than the combining them to form two features - "substituted by 1 word (S_1)" and "substituted by more than 1 words (S_m)".

We can further combine disfluencies (D) and substituted by more than one words (S_m) since the disfluencies can also be considered as "substituted by more than one word" in the process of self-correction, and the number of occurrences of these two categories is relatively low. Based on our observation that substitutions are mainly grapheme-to-phoneme pronunciation errors prevalent with Indian language speakers learning English, we further combine the fraction of correct words and words that are substituted by one word. For example, the words "heard" and "herd" are pronounced similarly, as /hɝd/. However, a student, who has never been taught the pronunciation of these words, can pronounce "heard" as /hɪəˈrd/. Such cases are not a fair representation of the student's reading skills, and hence, we can combine S_1 with C instead of penalizing the student. The three different input feature vectors used are also summarised in table 3.1.

Table 3.1: Input features considered for Lexical Clustering

Input feature vector	Description
4-dimensional (A)	$C + S_1$, M, I, $D + S_m$
5-dimensional (B)	C, S_1, M, I, $D + S_m$
6-dimensional (C)	C, S_1, S_m, M, I, D

3.2 K-means Clustering

To categorize the students based on their reading proficiency level, we have used an unsupervised clustering algorithm called K-means clustering. We have used the python implementation in the sklearn library [21].

K-means clustering aims at clustering the data into K clusters. The mean of the data points in each class, also called the cluster center, is taken as the prototype of the cluster. The clustering algorithm is given below :

1. Choose any K points randomly in the data-space. These are the initial cluster centers.

2. To assign a data-point, say P, to a cluster, take the Euclidean distance of P from each cluster center and assign P to the cluster corresponding to the nearest cluster center. Categorize all the data-points to get the K clusters.

3. For each cluster, find the mean of all the data points in the cluster. These are the new cluster centers.

4. If the updated cluster centers in step 3 are different from the centers at the beginning of step 2, perform step 2 again. Otherwise, the clustering is finished, and the required K clusters are those obtained at the end of the last iteration of step 3.

To remove the effect of randomly initializing cluster centers in step 1, this whole clustering is performed ten times, and the best clustering among these is taken as the final clustering. Facility for performing this repetitive clustering is present in the used sklearn implementation, where the best clustering is the best output of the ten runs in terms of inertia, i.e., the within-cluster sum-of-squares.

3.3 Evaluation Metrics

Silhouette score is a very commonly used evaluation metric for unsupervised clustering [22]. It captures the closeness of each sample to its own cluster compared to the other clusters. The score is a measure of how tightly grouped the clusters are, with a higher value indicating tighter grouping.

This score is computed for each data-point. Thus, for a new test data-point, it indicates the quality of clustering of that particular instance. It can be averaged over each class, giving us an indication of the relative quality of clustering of each cluster. It can further be averaged over all the data-points, used here, to give us an indication of the quality of clustering of the entire data-set.

For any data-point i, belonging to the cluster C_i, the intra-cluster variation is captured by a metric $a(i)$ and the inter-class variation is captured by a metric $b(i)$. These two are then used to compute the silhouette score $s(i)$ as explained below. Here, $d(i,j)$ denotes the euclidean distance between the data points i and j.

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(i,j)$$

$$b(i) = min_{i \neq j} \frac{1}{|C_j|} \sum_{j \in C_j} d(i,j)$$

$$s(i) = \begin{cases} \dfrac{b(i) - a(i)}{max\{a(i), b(i)\}} & \text{if} |C_i| > 1 \\ 0 & \text{if} |C_i| = 1 \end{cases}$$

$s(i)$ is then averaged over all the data-points in the dataset, to get the silhouette score of the entire dataset under the considered clustering.

$$\text{silhouette score} = \frac{\sum_i s(i)}{\sum_i 1}$$

The silhouette score ranges from -1.0 to 1.0. An analysis of different clustering scenarios and the corresponding silhouette values reveals that higher score, ideally 1, indicates a better clustering.

3.4 Results and Discussion

As mentioned above, we perform clustering using the different feature vectors mentioned in table 3.1. We also vary the number of categories from 2 to 6. The most appropriate cluster configuration is selected based on the interpretability of the clusters and the clustering quality, assessed by the silhouette score.

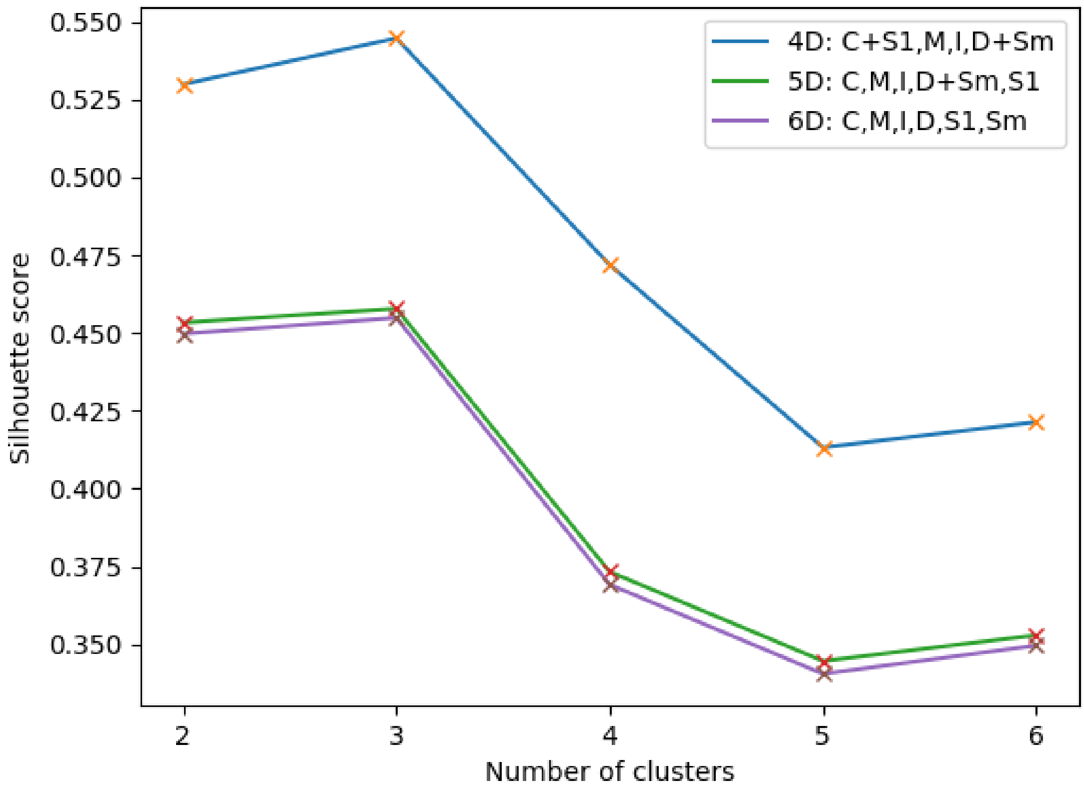


Figure 3.1: Average silhouette score versus number of clusters for the different feature vectors considered. Result when an entire story recording is considered an utterance.

We first consider an "utterance" to correspond to an entire story recording. The figure 3.1 shows the silhouette score for the different clustering configurations considered. The performance on using the 4D feature vector is significantly better than that using the other two feature vectors, 5D and 6D. This performance seems consistent with our observation that substitutions are mainly linked to valid attempts by the child that happen to be mispronunciations. We consider only the 4D feature in our further usage (i.e., classification using acoustic features, 5) and discussions as it performs the best among all others.

The obtained clusters in a 2-dimensional space of chosen 4D feature subset are shown in figure 3.2. Cluster 0, denoted by purple color code, corresponds to speech that is predominantly correct (or substituted by a single word). We call this class, C_A. The remaining two categories correspond to speakers with a lower proportion of C+S1 words. Cluster 1, denoted by blue color, is dominated by deletions or missing words and the cluster 2, denoted by the yellow color code, is dominated by gibberish or incorrect speech. We call these classes M_A and I_A, respectively. This observation provides the interesting insight that children with weak word-decoding skills do not necessarily pause and struggle to de-

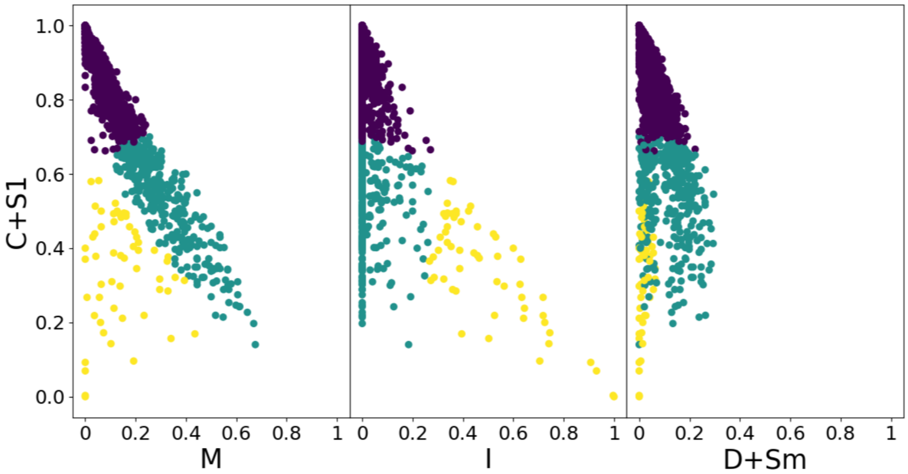


Figure 3.2: Scatter plots showing 3 clusters obtained using the 4D feature when the entire story recording in considered as an utterance.

code the individual words but can instead skim over the words while speaking a stream of gibberish. The latter scenario would be expected to be challenging to discriminate phonetically from correct speech without a powerful ASR system with a relatively unconstrained language model. This interpretation along with the number of instances categorized in each class is mentioned in Table 3.2. Due to the data available, the instances in the incorrect (gibberish) class are considerably low.

Table 3.2: Centers of the three clusters obtained using 4D input feature vector. An utterance is considered as the entire story recording.

Cluster	Color code	Interpretation	number of instances
CA	Purple	Predominantly correct words	687
MA	Blue	Predominantly missed words	329
IA	Yellow	Predominantly incorrect (gibberish) words	56

This interpretation of the clusters is also confirmed by the cluster centers mentioned in Table 3.3. C_A has a significantly higher fraction of (C+S1) words compared to the other two clusters. Hence, we say that this class represents the students who mainly speak correctly. Similarly, M_A and I_A have significantly higher fractions of missed and incorrect words, respectively, compared to the other two classes.

Apart from the silhouette score, the Euclidean distance between the cluster centers can also help us in deciding the number of classes. For example, if the distance between a pair of class centers is very less compared to that between any other pair, it is possible to club them together, i.e., a smaller number of total clusters would be better suited to the data. Table 3.4 shows the inter-cluster distance for the clustering into three classes using

Table 3.3: Centers of the three clusters obtained using 4D input feature vector. An utterance is considered as the entire story recording.

Cluster-feature	$(C{+}S_1)$	M	I	$D + S_m$
0	**0.85**	0.07	0.02	0.06
1	0.53	**0.31**	0.03	0.12
2	0.34	0.14	**0.49**	0.02

a 4D feature vector. The distance between the cluster pairs 0-1 and 1-2 is comparable, and that between the cluster pair 0-2 is slightly more but still comparable. Hence, the inter-cluster distances do not indicate a reduction in the number of clusters in our case.

Table 3.4: Euclidean distance between centers of the three clusters in Table 3.3

Cluster pair	Distance
0-1	0.40
1-2	0.54
0-2	0.69

Because of the variability in the reading styles and reading proficiency of the school children, the boundaries between the three classes need not be very strict. For example, in many cases, an utterance having 70 % correct and 30 % missed words has also been clustered into M_A, although one might expect it to be clustered in C_A. This is clearly a result of the data present and the clustering algorithm used. To deal with such cases, we cluster into more number of classes, thus giving us some intermediate classes, instead of just the three extremes seen here. Based on the classes observed and the application at hand (for example, if there is a requirement on the fraction of correct, missed or incorrect words for an utterance to be in any particular class), we can then interpret or combine classes. Here, as this categorization is simply an initial screening of the students before performing a fine-grained assessment of phrasing, prominence, etc., we can keep students having 70 % correct and 30 % missed in the C_A class. Although, utterance having 70 % correct and 30 % incorrect words can be kept in the I_A class because such utterances would give erroneous results when an ASR is used for further evaluation. An observation of the maximum percentage of incorrect words where we obtain acceptable results from the ASR can help to tune such thresholds.

In Figure 3.2, we see a large spread in the classes, especially in M_A and I_A. So, even though M_A can be said to primarily constitute of students having the highest fraction of missed words, many of them also have a significantly high fraction of correct words $(C + S_1)$. This also reflects from the cluster centers. The center of M_A has a higher

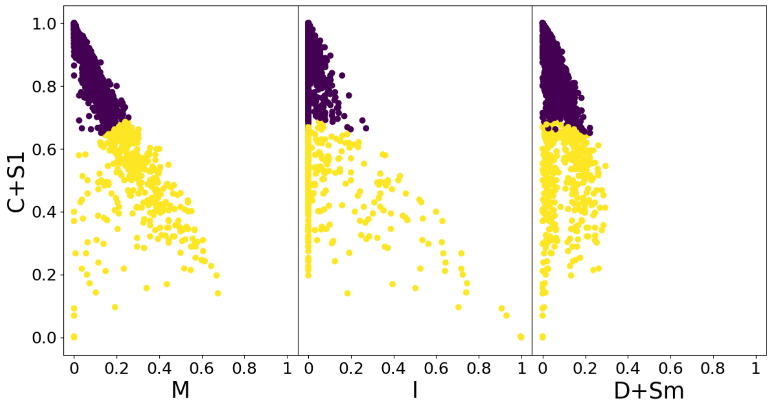


Figure 3.3: Scatter plots showing 2 clusters obtained using the 4D feature. An utterance corresponds to the entire story recording.

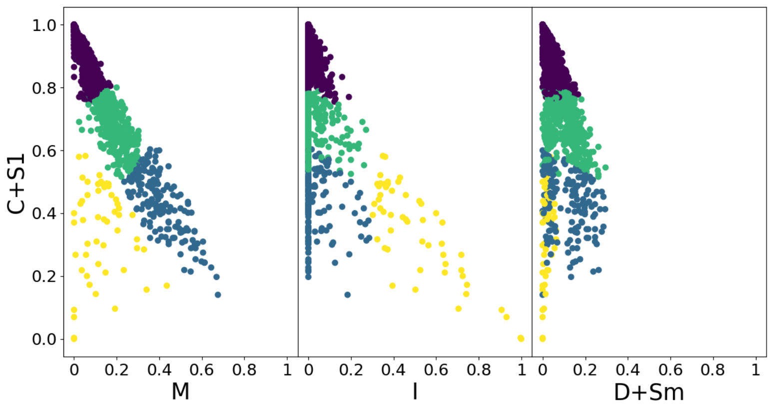


Figure 3.4: Scatter plots showing 4 clusters obtained using the 4D feature. An utterance corresponds to the entire story recording.

fraction of $(C + S_1)$ compared to that of M. This also motivates moving to a higher number of classes, as explained below.

Table 3.5: Euclidean distance between the cluster centers of four clusters obtained using 4D input feature vector, as shown in figure 3.4.

Cluster	0	1	2	3
0	0	0.59	0.26	0.74
1	0.59	0	0.33	0.56
2	0.26	0.33	0	0.60
3	0.74	0.56	0.60	0

Let us look into the interpretability of the clusters obtained using the 4D feature vector but with the number of clusters other than 3. Figure 3.3 shows that clustering into two classes gives a class of good speakers and another of struggling speakers. The latter category mainly comprises of the M_A and I_A in our 3-cluster clustering discussed above. Clustering into 4 classes, shown in Figure 3.4 gives an intermediate class apart from those

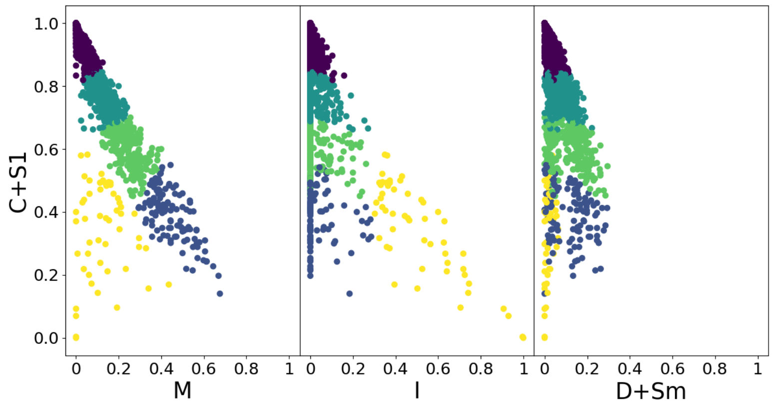


Figure 3.5: Scatter plots showing 5 clusters obtained using the 4D feature. An utterance corresponds to the entire story recording.

having the highest fraction of (C+S1), of M or of I. Since, this class has a higher fraction of (C+S1) compared to that of the other lexical features, i.e., M, I or D+Sm, we can combine it with the class of good, i.e., correctly speaking students (indicated by purple data-points, class 0). This is also supported by the Euclidean distance seen between the 4 cluster centers as mentioned in the table 3.5.

Similarly, we get two intermediate clusters when the clustering into five classes, and three intermediate classes when the clustering into six classes is performed. These clusters are shown in figures 3.5 and 3.6. In both these cases, combining the intermediate class with any of the predominantly correct, predominantly missed or predominantly incorrect classes, gives a very similar result as the original C_A, M_A and I_A classes in the 3-classes clustering. So, using the combination for classification is not very helpful. However, still performing a 5-way or 6-way classification can help us as we would be absolutely sure about the utterances classified in the extreme cases of correct (purple), missed (dark blue) and incorrect (yellow) classes.

A further examination of these classes reveals that the student's reading proficiency need not remain constant during an entire utterance. i.e., the student can have certain sections of the text with proper word decoding and can have very much trouble in word decoding in certain other sections. This could be because of the difficulty level of the text and can also be due to the student's skill level. Hence, performing this assessment at a smaller time-scale would provide more detailed feedback compared to comment for the whole recording. With our data, the smallest time-scale for which we have the manual lexical annotations is the individual sentences, where "sentences" are defined as the individual lines appearing on the Sensibol GUI shown in figure 2.2. However, considering each

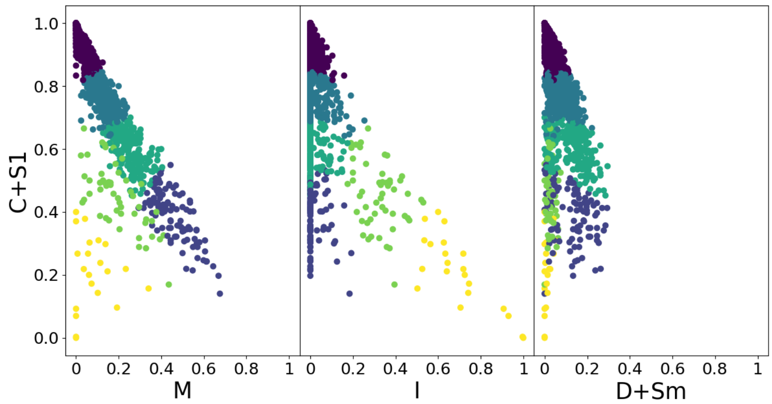


Figure 3.6: Scatter plots showing 6 clusters obtained using the 4D feature. An utterance corresponds to the entire story recording.

sentence as an utterance results in extremely short time intervals which are not sufficient to meaningfully evaluate the reading proficiency. So, we consider sentence-groups, i.e, chunks of sentences consisting of 20-30 words as an utterance. The different metrics of the number of words metrics of the number of words per utterance and the number of utterances is present in the table 3.6.

Table 3.6: Data distribution when clustering is performed at the sentence group level

Story	# words per utterance	# utterances per story	Total # utterances
The moon and the cap	27 - 30	5	1562
The lion and the fox	26 - 35	10	1347
The flying elephant	24 - 37	16	1248
The musical donkey	20, 28 - 35	8	727
Bunty and bubbly	22 - 29	5	1766
The wind and the sun	24 - 36	15	736

The results are seen similar to that in the story-level utterance scenario. The figure 3.7 shows the silhouette score for the different cluster configurations considered. As before, the performance on using the 4D feature vector is the best among the other two feature vectors, 5D and 6D, and so, we consider only the 4D feature in our further usage (i.e., classification using acoustic features, 5) and discussions.

Here, we see similar interpretation of the clusters as seen in the story-level case. Figure 3.8 shows the obtained clusters in 2-dimensional space of chosen feature subsets. We can interpret the three classes emerging here as follows. Cluster 0, denoted by purple color code, corresponds to speech that is predominantly correct (or substituted by a single word). We call this class "C_A". Cluster 2, denoted by yellow color code, represents

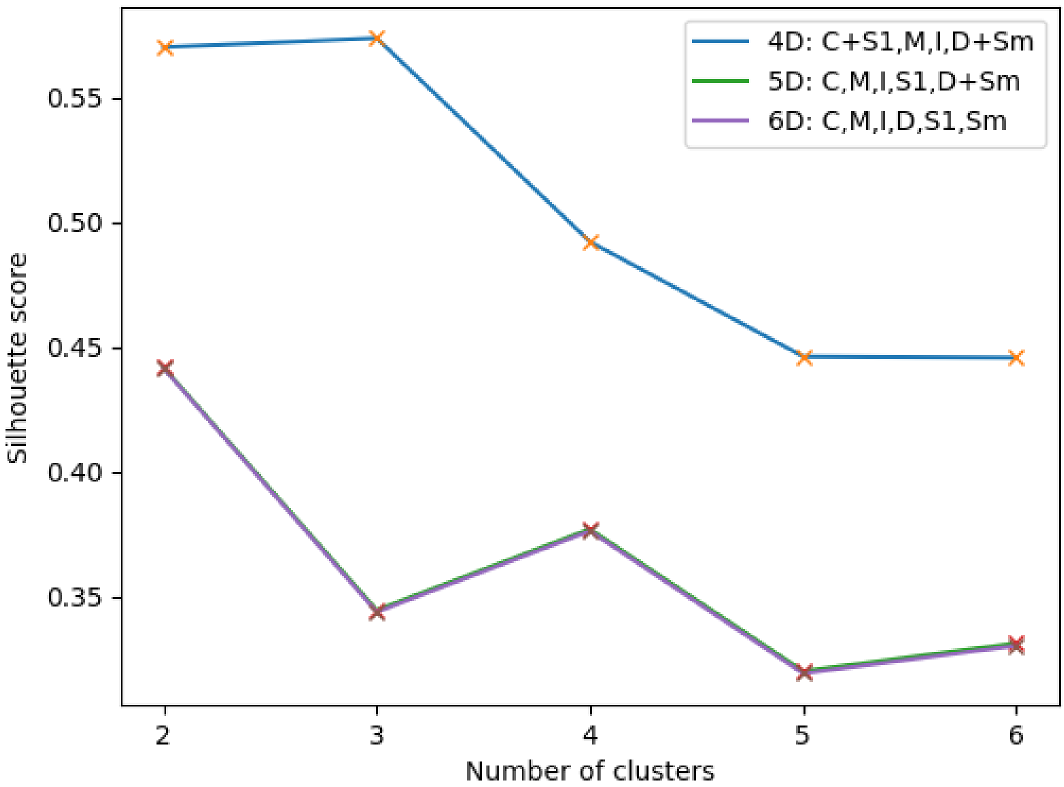


Figure 3.7: Average silhouette score versus number of clusters for the different feature vectors considered.

utterances dominated by gibberish speech and cluster 1, denoted by blue color code, represents utterances dominated by deletions or missing words. Here again, we call these classes "I_A" and "M_A" respectively. These interpretations of the three classes is confirmed by the centers locations as shown in table 3.8. This interpretation along with the number of instances categorized in each class is mentioned in Table 3.7. Here also, the number of instances in the incorrect (gibberish) class are considerably low due to the data available. The euclidean distance between the centers is shown in table 3.9.

By the same motivation as before, we examine the clustering into number of clusters different from 3. The results obtained are very similar to those obtained when an utterance was taken to be the entire story recording. Figure 3.9 shows the clusters when the clustering is performed into 2,4,5 and 6 classes respectively. All the interpretations and observations from the story-level scenario hold here as well. When clustered into 2 classes, we obtain a set of good speakers, almost like the C_A class when clustering into 3 clusters is performed and a class of poor speakers, which seems to be consisting of M_A and I_A classes from the 3-way clustering. When clustering is done into 4, 5 or 6 classes, we

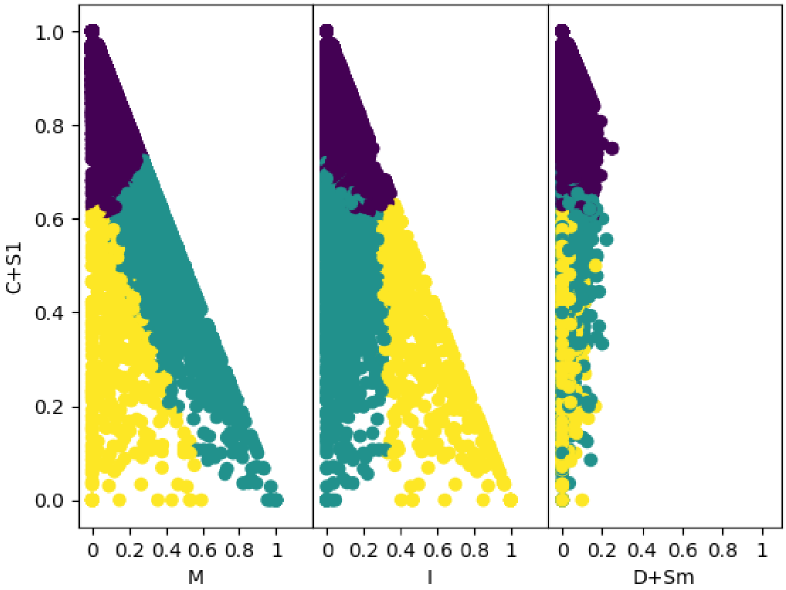


Figure 3.8: Scatter plots showing 3 clusters obtained using the 4D feature.

Table 3.7: Centers of the three clusters obtained using 4D input feature vector. An utterance is considered as the group of sentences.

Cluster	Color code	Interpretation	number of instances
C_A	Purple	Predominantly correct words	5291
M_A	Blue	Predominantly missed words	1799
I_A	Yellow	Predominantly incorrect (gibberish) words	449

Table 3.8: Centers of the three clusters obtained using 4D input feature vector

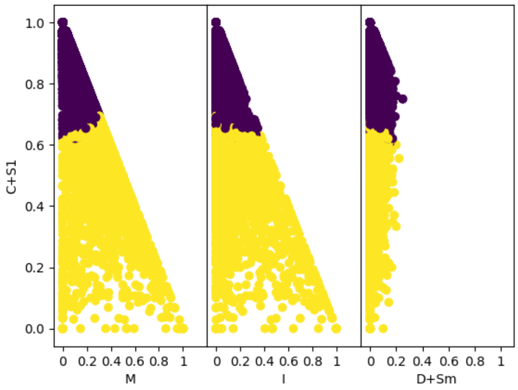
Cluster-feature	$(C+S_1)$	M	I	$D+S_m$
0	**0.88**	0.06	0.03	0.03
1	0.51	**0.38**	0.08	0.03
2	0.32	0.13	**0.54**	0.01

Table 3.9: Euclidean distance between centers of the three clusters in Table 3.8

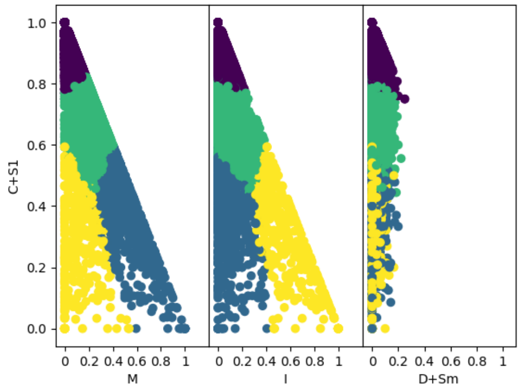
Cluster pair	Distance
0-1	0.49
1-2	0.56
0-2	0.75

obtain intermediate classes apart from the extreme cases of predominantly correct, predominantly missed and predominantly incorrect (shown by purple, dark blue and yellow color respectively in the respective figures).

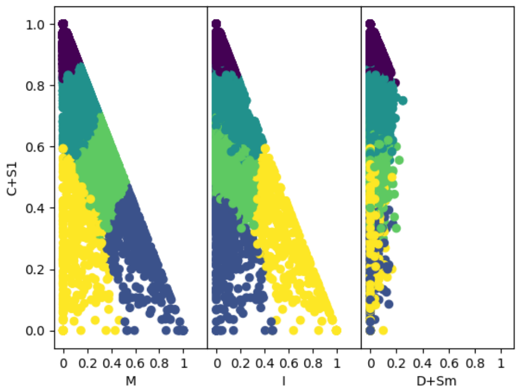
We next investigate acoustic features, easily computed from a recorded child-story in-



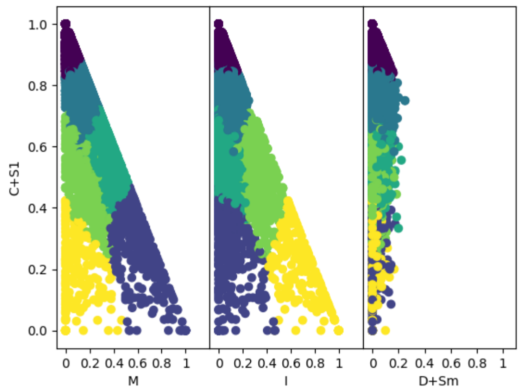
(a) 2 clusters



(b) 4 clusters



(c) 5 clusters



(d) 6 clusters

Figure 3.9: Scatter plots showing 2, 4, 5 and 6 clusters obtained using the 4D feature.

stance, that can provide cues for the three distinct lexical skill categories discovered from the manual transcriptions.

Chapter 4

Acoustic Feature Extraction

As mentioned in the previous chapter, we aim to categorize students based on their English reading proficiency without the use of ASR, using only the acoustic features derived from the audio. Features capturing pause behavior, syllable rate, and supra-segmental variations such as intensity and spectral dynamics have been observed to be useful for the task at hand. Table 4.1 lists these underlying attributes along with the specific low-level features considered. The following sections discuss these features, along with their importance and extraction.

Table 4.1: Description of extracted acoustic features

Attribute	**Feature**
Pause	mean, standard deviation (std), minimum, maximum of pause duration, pause freq, # pauses per video frame
Syllable rate (SR)	mean, std of relative # syllables, ratio of std and mean, articulation rate (AR)
Dynamics	Spectral centroid based (sp-dyn): FDR, NMC, NMV Intensity based (int-dyn): Macro-level variations and micro-level fluctuations

4.1 Pause-based features

An indicator of poor word decoding by the student is a higher fraction of pauses in the utterance. We detect silence regions using a VAD which is a combination of an Adaptive Linear Energy Thresholding based VAD and a ZFF based VAD [23]. Pauses are then defined as silences that exceed 200 ms. We compute the minimum, maximum, mean, and standard deviation of pause duration across the utterance. We also find the pause

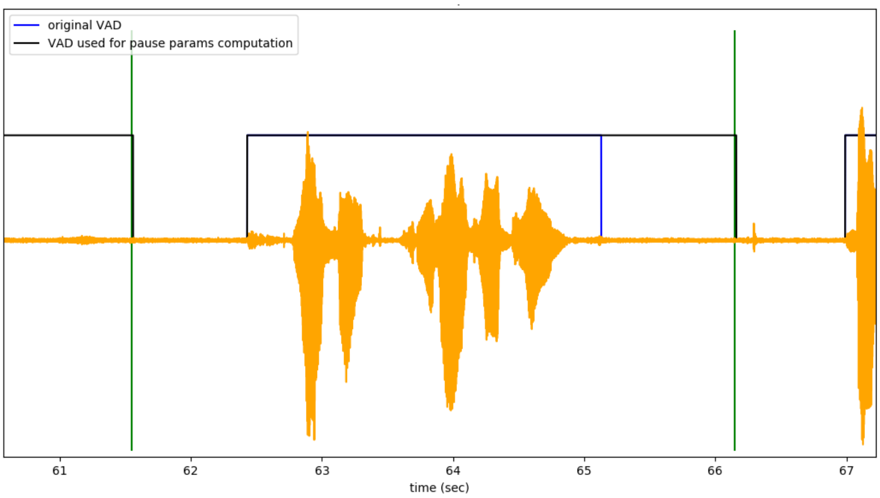


Figure 4.1: Text spoken:“One night she had a dream”. The vertical green lines are the time stamps where the video screen changes. The blue line shows the VAD decisions. The black line shows the modified VAD decisions used for our computations.

frequency, defined as the number of pauses per second, i.e., the total number of pauses divided by the total utterance duration. We also compute a feature, the pause frequency in a video frame, i.e., the number of pauses in the utterance divided by the number of video frames in the utterance.

A higher silence region in the audio can be a result of a struggle in word-decoding. Because our recording system is video-karaoke based, it can also be because the student quickly read the text given on the video screen and is now waiting for the screen to change to read the other part of the text. To not penalize the student if the second reason the case, we have taken a simple approach to ignore a pause just before the video change boundary completely. If a video change boundary lies in a pause region, we ignore the pause just before the boundary and the pause after the boundary is taken unchanged. Figure 4.1 illustrates the same. Another approach could be to reduce only a fixed amount from the pause before the video change boundary, but deciding this fixed amount is not easy given the variability in our data.

4.2 Syllable rate based features

The syllables spoken by the student are important indicators of his/her proficiency level as described below. Note that here we are merely interested in the syllables spoken, not the “correctly” spoken syllables as we are not using an ASR. We compute the relative number of syllables in each video interval, which is defined as the fraction of detected syllables in the video interval to the number of syllables in the known text corresponding to the

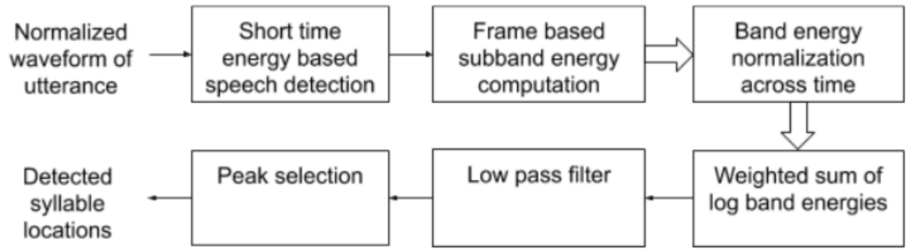


Figure 4.2: System block diagram for syllable detection. [1]

same interval. The relative number of syllables in a video frame should ideally be 1. As the deviations increase on either side of 1, the proficiency level of the speaker can be said to be decreasing. Ideally, we would expect a student struggling in word-decoding to have the relative number of syllables less than 1. However, due to the "gibberish" phenomena seen in our data, the relative number of syllables can be high even for struggling speakers. In some cases, this ratio is even higher than 1 for a gibberish-speaking student. As mentioned in section 3.1, substitutions by one word are combined with correct words for all the experiments. So, the relative number of syllables can be greater than 1 even for the correct class, C_A. This ratio is expected to be low for the class with predominantly deletions, M_A.

The mean and standard deviation of the relative number of syllables are computed across the video intervals corresponding to the utterance. We also compute the overall articulation rate as the total number of syllables detected by the total speech duration of the utterance.

As every syllable would have a vowel at its nucleus, we approximate the number of syllables by the number of vowels. Now, for the accurate estimation of the number of vowels, we estimate a representative weighted sub-band energy contour having peaks at vowel locations. This system is shown in the figure 4.2. More details on this system can be found in [1].

4.3 Spectral Centroid based features

On observing the data, we see that the gibberish speakers often repeat the phones spoken. Because the phones, and so, the formants are repeated, we see a repetitive structure in the frequency distribution of the audio. We use the spectral centroid to capture this. As the name suggests, the spectral centroid is a form of average over the frequencies present in the audio. The python implementation in the Librosa library is used [24]. We compute spectral centroid over the entire 0 - 8 kHz band from the short-time spectrum every 10 ms frame. It is quantized to 400 Hz bands. Silence regions are discarded based on VAD decisions. We have also observed that spectral centroid lying in the region 2500 Hz to 3500 Hz, corresponds to silence or noise regions. So, these regions are also discarded. For the remaining speech region in a given video interval, we count the most frequently occurring spectral centroid band and the second most frequently occurring band, c1, and c2 respectively.

Using c1 and c2, we compute the following features :

- Frequency distribution ratio (FDR): Ratio close to unity indicates energy distribution across the frequency range. Higher the ratio, higher is the single frequency band dominance in the recording, a characteristic of incorrect speech where articulation is relatively repetitive in nature.

$$FDR = \frac{c_1}{c_2}$$

- Normalized mode count (NMC): It captures the dominance of the most frequently occurring spectral centroid band in the speech regions of the utterance. As c1 would be higher for incorrect (gibberish) speech compared to correct speech, this feature is expected to be higher for the incorrect speech.

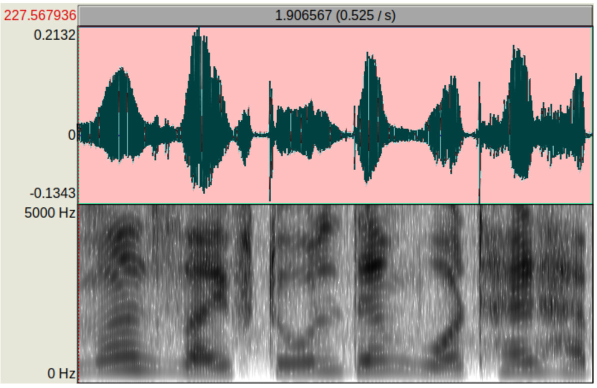
$$NMC = \frac{c_1}{\text{speech duration in utterance}}$$

- Normalized mode variation (NMV): It measures the variation in the dominance of highest occurring frequency in each video frame across the audio. This feature is more useful for utterances at a larger time-scale, for example, at the story level, as measures the consistency of the speaker over different video intervals. When the length of the utterance reduces, this feature does not have much meaning. This is also seen from the experimental results as mentioned in the section **??**.

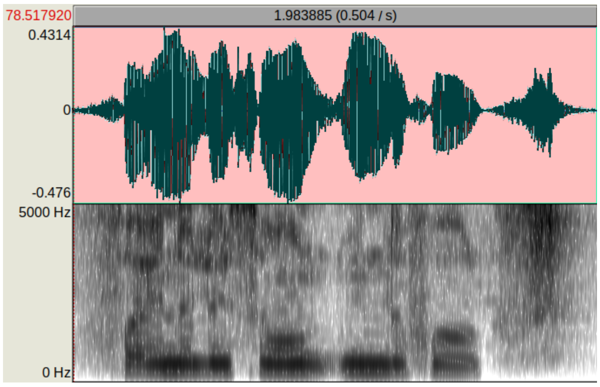
$$NMV = std(\frac{c_1}{\text{speech duration in utterance}})$$

4.4 Intensity based features

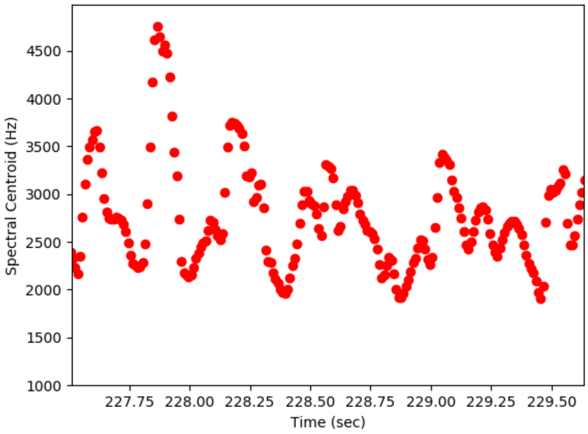
In the case of unintelligible speech, the dynamics are less prominent due to mumbling. The intensity contours are more fluctuating in correct students and smoother in gibberish



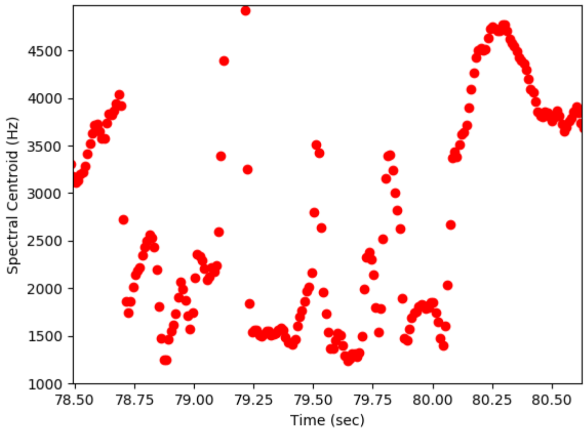
(a) Spectrogram for an utterance in C_A.



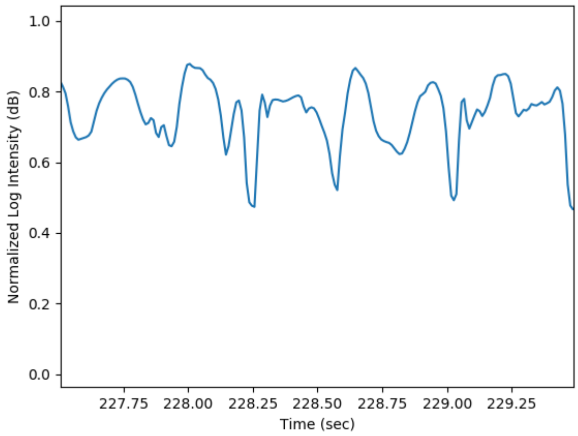
(b) Spectrogram for an utterance in I_A.



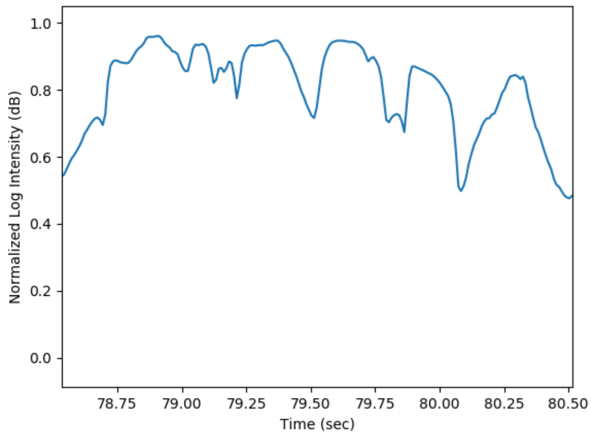
(c) Spectral Centroid for an utterance in C_A class.



(d) Spectral Centroid for an utterance in I_A class.



(e) Normalized Log Intensity contour for an utterance in C_A class.



(f) Normalized Log Intensity contour for an utterance in I_A class.

Figure 4.3: Spectrogram, spectral centroid and log normalized intensity contours for an utterance in class C_A, shown in (a), (c) and (d); and for an utterance in class I_A, shown in (b), (d) and (f). Text spoken in the utterance in C_A: "wife told them what had". Text spoken in the utterance in I_A: "finally sond inda fon" (actual text: finally she found a man)

speaking students. These inter-syllable and intra-syllable variations have been illustrated in Figures 4.3e and 4.3f for an utterance in correct class and one in incorrect class respectively. The features computed based on these observations are mentioned below. The short time intensity contour computed across 10 ms frames exhibits different kinds of dynamics at small and larger (i.e., across syllables) time scales. Intensity normalization is performed at the video frame level followed by VAD based silence removal. We then compute the following features :

- Intensity contour smoothness at syllable level: The intensity contour is averaged over 300 ms windows. For each video interval, we compute the standard deviation of the contour. Mean and standard deviation of this feature across the recording capture the inter-syllable variation in intensity. This variation is less for the incorrect speech compared to the correct speech, so we expect the mean to be low for incorrect speech.

- Intensity contour smoothness at 10-40 ms level: We compute measures of the fluctuations as the mean and standard deviation of the contour fluctuations in 50 ms time windows. These contour fluctuations are found by averaging auto-correlation of the short term energy computed over frame size of 10ms with lag up to 5 frames. This averaged auto-correlation (ACR_{avg}) is given by:

$$ACR_{avg} = \frac{1}{lag} \sum_{k=0}^{lag-1} \sum_{n=0}^{\infty} STE[n] \cdot STE[n+k],$$

  where $STE[n]$ is short term energy for n^{th} frame and padded with zeros wherever necessary while computing the above ACR_{avg}. These are expected to measure the speaker's intensity variations at a micro level. For smooth intensity contour, the autocorrelation would be high and so, the mean of the averaged autocorrelation is expected to be high for incorrect speech.

Chapter 5

Classification Results and Discussion

The acoustic features extracted above have been used to classify the students into three classes - containing predominantly correct words 'C_A', predominantly missed words 'M_A' and predominantly incorrect (gibberish) words 'I_A'. We have used supervised classification for this with the ground-truth training labels obtained from clustering. Random Forest classifier, implemented in the sklearn library of Python, has been used for the classification. We first perform the classification considering the entire story recording as an utterance. However, as mentioned in section 3.4, the student's performance might be different in different parts of the story. We also look into the results of the classification when sentence-groups are taken as utterances.

First, a single stage classifier is tested with the extracted features as input vector per utterance. As some of the acoustics features were observed to be efficient in distinguishing I_A, we use a 2-stage classifier (P). It is expected to separate I_A from the other two classes in the first stage, and C_A and M_A in the second stage. In our data, the C_A and I_A classes contain mostly speech as opposed to the M_A class. This motivates another 2-stage classifier (Q) to separate M_A from C_A and I_A first, and then separate C_A and I_A. The ground truth labels of the C_A, M_A and I_A classes are obtained from the lexical clustering as mentioned in section 3.4. The performance is evaluated using accuracy score, i.e., the percentage of correctly classified instances among the total data-points present.

5.1 Random Forest Classifier

A random forest classifier [25] is essentially an ensemble of multiple decision trees, where each tree is trained, and the final output is the mode of the outputs of each of the decision trees. This merging of the output from multiple trees provides a more accurate and stable prediction. A decision tree classifier is exactly what is suggested by the name: a tree-like decision-making model with a node representing a decision on the inputs, a branch representing the outcome of that node's decision and a final node with no branches, leaf node, representing the output of the decision tree. To decide which feature to use while splitting a node, the decision tree classifier uses the feature which contributes the most in distinguishing the classes. Parameters like Gini impurity are used for this evaluating the importance of a feature.

Apart from the forest created by an ensemble of the decision trees, a random forest adds randomness to the model by considering only a random subset of features for splitting a node, instead of the entire feature space as in a decision tree. This results in diversity across the trees, generally resulting in a better model. Reducing the feature space and building smaller trees also prevents overfitting.

Another quality of the random forest classifier is that we can find the relative importance of each feature for the prediction. The sklearn implementation used, measures the feature importance by looking at how much the tree nodes, which use that feature, reduce impurity across all trees in the forest. This score is computed for all the features after training, and the computed feature importances are scaled so that their sum is 1. The feature importances can help in selecting which features to use and which to drop from the classification.

Because of the good prediction result and the ease of understanding and using the hyperparameters, random forest classifier is considered a handy tool. With enough trees, overfitting is also prevented.

5.2 Results

This section first shows the experimental results when story-level utterances are used. The results using smaller sentence-group level utterances are presented next. For both the cases, we first perform a single stage classification into the three classes, C_A, M_A and I_A. As the dynamics features are observed to be efficient in detecting 'incorrect' speech, we also test a 2-stage classifier (P). It is expected to separate I_A from C_A and M_A in the first stage and to the separate C_A and M_A in the second stage. In our data, the C_A and I_A classes contain mostly speech as opposed to the M_A class. This motivates another 2-stage classifier (Q) to separate M_A from C_A and I_A first, and then separate C_A and I_A. For the three classifier systems used, we have tried out different input feature combinations and the results with the best performing classifiers are shown below.

We can also classify the students into two categories, viz., students with good word decoding, those that were classified into C_A, and students with poor word decoding, those that were classified into M_A or I_A in our 3-way clustering. This is useful for us because we can then perform further prosodic assessment for the students in the C_A category using an ASR, but using and ASR for utterances in M_A and I_A would give erroneous results. This classification can be done is a single stage. It can also be done using the results from the 3 way classification above. Using the best performing 3-way classifier, predictions into classes M_A and I_A can be combined as they have been considered the same class. Note that we are not re-training the classifier in this case. The following subsections contain these results for both sentence-level utterances and for smaller sentence-group level utterances.

5.2.1 Results for story-level utterances

The results when the entire story is considered as an utterance are presented first. 7-fold cross validation has been used. Note that the data used for classification is kept unbiased towards any class. In our data, we have the least number of instances for the incorrect class. So, the data used for classification is limited by the number of instances in this class. At the story-level, the instances in the three classes as mentioned in Table 3.2 have been considered. All the 56 instances in the incorrect class have been considered for classification. 63 utterances corresponding to the missed class and 70 utterances corresponding to the correct class have been considered. For the three classifiers, the best performing configurations and their results are mentioned in Table 5.1.

The confusion matrix of the best performing scheme (P) in the Table 5.1 is shown in the Table 5.2. The misclassifications are found to be more in the missed class compared to the others; this is because whenever the students in the missed class spoke, they were correct (largely) or incorrect. We can say that the missed class bridges the gap between correct and incorrect. This can also be observed from the cluster plots in Figure 3.2. We have

Table 5.1: Classifiers and obtained accuracies in 3-way classification. An utterance refers to the entire story recording.

Classifier Configuration	**Feature Combination**	**Accuracy**
1-stage	Pause features, SR features, sp-dyn, int-dyn	65.7 %
2-stage (P)	Stage 1: AR, sp-dyn, int-dyn, #pauses per video frame Stage 2: pause features, SR features	68.3 %
2-stage (Q)	Stage 1: pause features, SR features, sp-dyn, int-dyn Stage 2: sp-dyn, int-dyn, SR features, pause freq	64.6 %

also observed that the correct class is more confined whereas the other two classes display varying amounts of their respective characteristics. So, we expect more confusion in classifying these. The dynamics features chosen, sp-dyn and int-dyn, are designed to pick the incorrect (gibberish) speaking students from the others, improving the classification of incorrect class. This motivated the two-stage classification (P). On the other hand, we were unable to find similar tailored features for the missed class. Another performance degrading parameter in our data is the presence of different kinds of noise, ranging from white noise to background talkers. The features expected to classify 'missed' recordings (pause features) are adversely affected by this. This explains the comparatively lower accuracies for the missed class in Table 5.2.

Table 5.2: Confusion matrix for the highest accuracy classifier (P) of Table 5.1

Actual \ Predicted	$\mathbf{C}_A$	$\mathbf{M}_A$	$\mathbf{I}_A$
$\mathbf{C}_A$	51	7	12
$\mathbf{M}_A$	15	35	13
$\mathbf{I}_A$	7	6	43

The accuracy seen in the single stage 2-way classification is 69.3 % with confusion matrix as shown in Table 5.3. Here, we have used 70 utterances for each of the two classes, the C_A class and students with poor word decoding with 35 utterances each for M_A and I_A. Performing the 2-way classification using results from the best performing 3 way classifier (P) above, gives 78.3 % accuracy with confusion matrix shown in Table 5.4.

Table 5.3: Confusion matrix for the single-stage 2-way classification.

Actual \ Predicted	$\mathbf{C}_A$	$\mathbf{M}_A + \mathbf{I}_A$
$\mathbf{C}_A$	49	21
$\mathbf{M}_A + \mathbf{I}_A$	22	48

Table 5.4: Confusion matrix for the 2-way classification using the best performing 3-way classifier (P).

Actual \ Predicted	$\mathbf{C}_A$	$\mathbf{M}_A + \mathbf{I}_A$
$\mathbf{C}_A$	51	19
$\mathbf{M}_A + \mathbf{I}_A$	22	97

5.2.2 Results for sentence-group-level utterances

The results considering a group of sentences as an utterance are presented now. As each story level utterance is divided into sentence-group-level utterances, more utterances in available for use. This data has also been mentioned in Table 3.6. 15-fold cross validation is used. The random forest used consists of 500 trees. Here also, the data used for classification is kept unbiased towards any class. As, we have the least number of instances for the incorrect class, the data used for classification is limited by the number of instances in this class. Out of the 449 instances in the I_A class, some utterances are removed due to very high noise and 435 utterances are used for the classification. 450 utterances of M_A and 465 utterances of C_A are used for the classification. The accuracies seen here, mentioned in Table 5.5, are slightly less than those seen at the story level.

Table 5.5: Classifiers and obtained accuracies in 3-way classification. An utterance refers to a group of sentences.

Classifier Configuration	**Feature Combination**	**Accuracy**
1-stage	All except ratio of std and avg rel num syl, NMV	62.88 %
2-stage (P)	Stage 1: all except min, max pause duration, NMV, # pauses / # video intervals Stage 2: all except sp-dyn, syllable level intensity in int-dyn	61.55 %
2-stage (Q)	Stage 1: all except NMV Stage 2: all except NMV	61.48 %

The confusion matrix for the best performing 1-stage classification has been shown in the Table 5.6. The accuracy seen in the single stage 2-way classification into students with good word-decoding, C_A and students with poor word-decoding, M_A and I_A is 72.2% and the confusion matrix in shown in Table 5.7. Here, 465 utterances for each of the two classes are used with 240 for M_A and 225 utterances for I_A. As done for story-level above, we can use the best performing 3-way classifier and combine the predictions of classes M_A and I_A as they have been considered the same class. Again, the classifier has not been retrained. The accuracy obtained is 74.1 % and the confusion matrix is shown in Table 5.8.

Table 5.6: Confusion matrix for the 1-stage classifier in Table 5.5

Actual \ Predicted	$\mathbf{C}_A$	$\mathbf{M}_A$	$\mathbf{I}_A$
$\mathbf{C}_A$	294	84	87
$\mathbf{M}_A$	93	278	79
$\mathbf{I}_A$	86	72	277

As mentioned before, the feature NMV measures the variation in the dominance of highest occurring frequency in each video frame across the audio. This was particularly useful when the utterance considered was at the story level. However, the smaller sentence-group-level utterance already captures these small scale variations and hence, this feature was not found to be very useful.

Table 5.7: Confusion matrix for the single-stage 2-way classification.

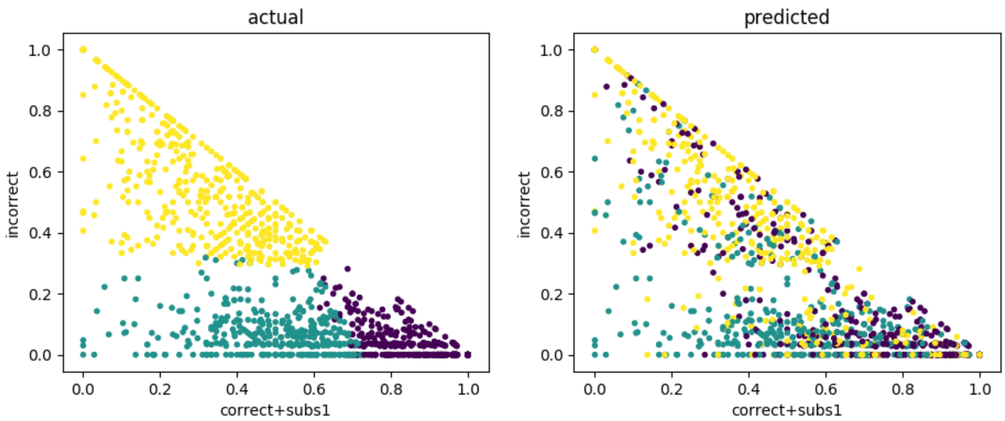
Actual \ Predicted	$\mathbf{C}_A$	$\mathbf{M}_A + \mathbf{I}_A$
$\mathbf{C}_A$	339	126
$\mathbf{M}_A + \mathbf{I}_A$	132	333

Table 5.8: Confusion matrix for the 2-way classification using the best performing 3-way classifier.

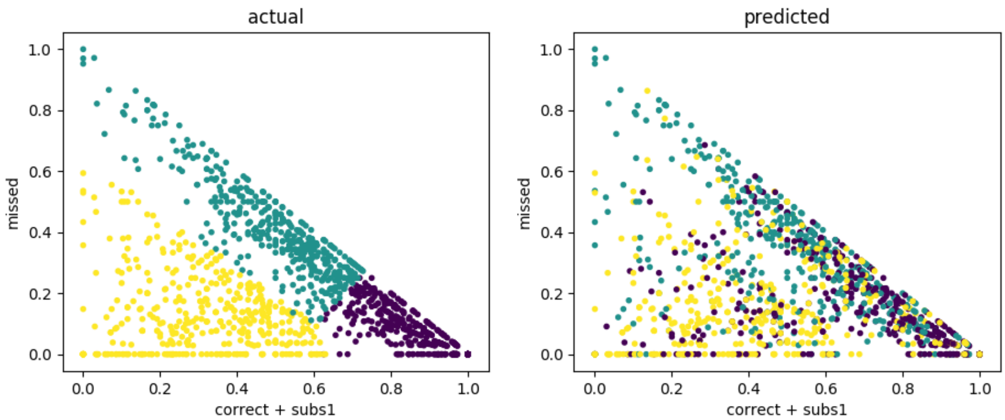
Actual \ Predicted	$\mathbf{C}_A$	$\mathbf{M}_A + \mathbf{I}_A$
$\mathbf{C}_A$	294	171
$\mathbf{M}_A + \mathbf{I}_A$	179	706

5.3 Discussion

This section contains a discussion on the results seen for the sentence-group level case. The same discussion follows for the story-level case. Figure 5.1 shows the actual and predicted classes in the lexical feature space. As can be seen, there is lesser confusion in the area towards the extremes in the three lexical features (M, I, and C+S1) compared



(a) I v/s (C+S1)



(b) M v/s (C+S1)

Figure 5.1: The actual and predicted classes are shown in the 2-dimensional lexical feature space. As in Figure 3.8, the Purple color code is used for the C_A class, the Blue color code is used for the M_A class and the Yellow color code is used for the I_A class.

to that in the middle. The challenges in our data and aspects of the features considered, which lead to errors in our results are discussed next.

Challenges in the data leading to errors in our categorization system:

- **Video-karaoke form of recording**: A significant challenge is the high amount of variability in our data, especially that introduced by the video karaoke form of recording. If the video screen changes and the student has still not finished speaking the text, he/she can continue speaking it from memory in the next video screen. In such a case, the lexical annotations need not match the acoustic characteristics of both the video frames. On the other hand, a good student might finish speaking the text before the video frame ends. In such a case, the pause before the video frame boundary would not be due to the incompetence of the student. We are unable to distinguish between such a pause and a pause just before the video frame where the student actually could not speak. As mentioned in section 4.1, we have removed this pause completely from our pause feature computations. But this has caused confusion in cases with pauses of the latter kind, i.e., where the pause just before a video frame boundary is actually due to the student being unable to speak the text. Due to this pause information being lost, some students in the M_A category have got misclassified into the C_A category.

- **Transcription system limitation**: A simple energy-based VAD was used while uploading the audios on the manual transcription interface. This VAD was later upgraded to a more sophisticated ALED and ZFF based VAD [23]. In many cases, some sections of the speech have also been removed by the VAD. These deducted parts of the speech have not been shown on the interface for the lexical transcriptions, and so, many words have been marked missed, even though the student has spoken them. In such cases, the utterance should actually be in the class C_A, and it is also classified as C_A, but because it has been marked M_A in the lexical clustering, it is counted as an error while computing the classifier accuracy.

- **Presence of noise**: As mentioned earlier, we have considered audios with low noise for the C_A and the M_A class, but as we already had significantly less number of audios for the I_A class, we had considered all of them. This noise in the incorrect audios leads to erroneous results from the feature computation blocks, which further leads to erroneous classification.

Aspects of the features considered that lead to errors in our categorization system:

- **Syllable detection block**: The syllable detection block used has been currently trained on clean data. As our data is noisy, it often gives more syllables than that are actually present, hence, accounting for errors, especially the misclassification of utterances in the class M_A into classes C_A or I_A.

- **Frequency dominance**: The spectral centroid computed is based on the premise that the students in I_A speak repetitively. But some of them have spoken in a very confident manner, along with modulations. For such cases, our dynamics based features are not able to distinguish such audios from those in the C_A class.

- **Micro-level intensity variation**: The feature importance obtained from the Random Forest classifier shows that the micro-level variations in the intensity play a significant role in the classification. Many correctly speaking students have been observed to have smooth intensity contours, i.e., very less micro-level intensity variations, which were primarily found to be characteristic of the speakers in I_A class. The cause for this is not completely clear yet.

Chapter 6

Conclusions and Future Work

6.1 Conclusion

Based on the fact that beginning (second-language) readers come with diverse skill levels in the word-decoding aspect of oral reading, we attempted to characterize the behaviour of the population represented in our children's reading data set. The best underlying clusters in lexical miscues space turned out to correspond to good readers and two types of lower-proficiency readers. It emerged that poor word decoders can not only skip words they cannot recognize, but also resort to speaking gibberish (unintelligible stream unconnected to the text). We proposed acoustic signal features to discriminate the incorrect (unintelligible) speech from correct speech, and overall, achieve the categorization of speakers into the 3 classes. Since the number of instances that were labeled incorrect was relatively small, we had to restrict our test data set.

On observing that the student might have good word-decoding for certain parts of the story but poor word decoding for certain other parts, we perform the analysis for smaller utterances, i.e., dividing each story recording into chunks, each consisting of 20-30 words. Reducing the time-scale also increased the data available to us by a significant amount. Analysing at the sentence level would also increase the confusion as even the smaller variations in the acoustic features, which were being averaged out earlier would play a significant role now. Also, we now have much more extensive dataset compared to what we had previously, adding to the variability in the training and testing data. The classification accuracy seen at this sentence-group level is comparable to that obtained at the story-level. Hence, with the added benefit of giving better feedback to the students and the added difficulty in the data, a similar accuracy is clearly acceptable. Using the sentence-group level classification, we can make better comments on the reading proficiency of the student. For example, we can say which fraction(s) of the story recording was spoken correctly, incorrectly, or primarily missed by the student.

6.2 Future Work

- We can perform these experiments using enhanced, i.e., denoised, speech. It was observed that both SEGAN-based speech enhancement and traditional signal processing based speech enhancement distort the pitch and intensity in the speech regions [26], [27]. Hence, extracting the acoustic features from the denoised speech might not be accurate, but it can be examined. Instead of using the denoised speech for acoustic feature extraction, we can use it for better speech-silence demarcation, i.e., better VAD decisions.

- The video karaoke form of recording introduces a lot of complications. Performing this classification on read text, but where the text is not presented in video karaoke form, could be very useful and much more insightful about the student's performance; independent of the effects introduced by the video-karaoke form of recording.

- Currently the data used in classification is limited by the data available for the I_A class. More data would give a better insight in the student's proficiency level. We would also be able to use advanced machine learning techniques such as Deep Neural Networks, Support Vector Machines, etc. More features can be introduced to improve the performance. For example, we have not made use of the fact that the speakers in the I_A class had spectral centroids towards the lower ends of the frequency spectrum.

- Probabilistic classification can be advantageous here because of the large variability of the data. I.e., instead of a verdict of which class the utterance belongs to, we can tell the probability of it being in the three classes.

- ASR output can be checked for the three classes obtained to verify that ASR performs poorly on the I_A and M_A classes compared to the C_A class.

List of Publications (Submitted)

1. C. Vitthal, K. Sabu, Shreeharsha B S, and P. Rao, "Predicting lexical skills from oral reading with acoustic measures", Interspeech, Graz, Austria, 2019.

References

[1] K. Sabu, S. Chaudhuri, and P. Rao. An optimized signal processing pipeline for syllable detection and speech rate estimation. In *submitted to INTERSPEECH*, 2019.

[2] J. Miller and P. Schwanenflugel. A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. *Reading Research Quarterly*, 43(4):336–354, 2008.

[3] M. Breen, L. Kaswer, J. Van Dyke, J. Krivokapic, and N. Landi. Imitated prosodic fluency predicts reading comprehension ability in good and poor high school readers. *Frontiers of Psychology*, 7:1–17, 2016.

[4] P. Schwanenflugel, A. Hamilton, J. Wisenbaker, M. Kuhn, and S. Stahl. Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers. *Journal of Educational Psychology*, 96(1):119–129, 2004.

[5] ASER: The Annual Status of Education Report (rural). `http://img.asercentre.org/docs/ASER%202018/Release%20Material/aserreport2018.pdf`. ASER Centre 2018, last accessed 9/6/2019.

[6] K. Sabu, P. Swarup, H. Tulsiani, and P. Rao. Automatic assessment of children's L2 reading for accuracy and fluency. In *Proceedings of SLaTE*, Stockholm, Sweden, 2017.

[7] K. Sabu and P. Rao. Detection of prominent words in oral reading by children. In *Proceedings of Speech Prosody*, Poznan, Poland, 2018.

[8] M. Black, J. Tepperman, S. Lee, P. Price, and S. Narayanan. Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment. In *Proceedings of INTERSPEECH*, Belgium, 2007.

[9] National Reading Panel. Teaching children to read:an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Technical report, The Eunice Kennedy Shriver National Institute of Child Health and Human Development, 2000.

[10] M. Black and S. Narayanan. Improvements in predicting children's overall reading ability by modeling variability in evaluators' subjective judgments. In *Proceedings of ICASSP*, Kyoto, Japan, 2012.

[11] X. Li, L. Deong, Y. Ju, and A. Acero. Automatic children's reading tutor on hand-held devices. In *Proceedings of INTERSPEECH*, Brisbane, Australia, 2008.

[12] J. Cheng, Y. D'Antilio, Chen, and J. Bernstein. Automatic assessment of the speech of young english learners. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, Baltimore, Maryland USA, 2014.

[13] K. Zechner, D. Higgins, X. Xi, and D. Williamson. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication*, 51(3):883–895, 2009.

[14] Pearson - Versant spoken language tests, patented speech processing technology, and custom test services. `www.versanttests.com`. Pearson Education Inc, 2016.

[15] J. Mostow. Why and how our automated reading tutor listens. In *Proceedings of International Symposium on Automatic Detection of Errors in Pronunciation Training*, Stockholm, Sweden, 2012.

[16] D. Bolanos, R. Cole, W. Ward, G. Tindal, P. Schwanenflugel, and M. Kuhn. Automatic assessment of expressive oral reading. *Speech Communication*, 55(2):221–236, 2013.

[17] J. Liscombe. *Prosody and Speaker State: Paralinguistics, Pragmatics, and Proficiency*. PhD thesis, Columbia University, 2007.

[18] L. Fontan, M. Le Coz, and S. Detey. Automatically measuring l2 speech fluency without the need of asr: A proof-of-concept study with japanese learners of french. In *Proceedings of INTERSPEECH*, Hyderabad, India, 2018.

[19] Bookbox: A book for every child in her language. `www.bookbox.com`. redirected to PlanetRead's website (a partner of BookBox) on 13 June 2019: `https://www.planetread.org/anibooks`.

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

[21] scikit-learn:mmachine learning in python. `http://scikit-learn.org/stable/`. Scikit-learn.

[22] P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, 20:53–65, 1987.

[23] A. Pasad, K. Sabu, and P. Rao. Voice activity detection for children's read speech recognition in noisy conditions. In *Proceedings of National Conference on Communications*, Chennai, India, 2017.

[24] Librosa:a python package for music and audio analysis. `http://scikit-learn.org/stable/`. Librosa.

[25] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[26] C. Vitthal. "Denoising of Speech Signals". *Supervised Research Exposition, Department of Electrical Engineering, IIT Bombay*, 2018.

[27] K. Sabu, K. Kumar, and P. Rao. "Improving the Noise Robustness of Prominence Detection for Children's Oral Reading Assessment". *Proc. of NCC, Feb, Hyderabad, India*, 2018.