Investigation into the Effect of Dynamic Acoustical Features in Deep Neural Network-based Speech Classification

by

Md. Rakibul Hasan

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science in Electrical and Electronic Engineering

Khulna University of Engineering & Technology

Khulna 920300, Bangladesh

October 2021

Declaration

This is to certify that the thesis work entitled *"Investigation into the Effect of Dynamic Acoustical Features in Deep Neural Network-based Speech Classification"* has been carried out by *Md. Rakibul Hasan* in the Department of *Electrical and Electronic Engineering*, Khulna University of Engineering & Technology, Khulna, Bangladesh. The above thesis work or any part of this work has not been submitted anywhere for the award of any degree or diploma.

Signature of Supervisor

Signature of Candidate

Approval

This is to certify that the thesis work submitted by *Md. Rakibul Hasan* entitled *"Investigation into the Effect of Dynamic Acoustical Features in Deep Neural Network-based Speech Classification"* has been approved by the board of examiners for the partial fulfillment of the requirements for the degree of *Master of Science in Engineering (M. Sc. Eng.)* in the Department of *Electrical and Electronic Engineering*, Khulna University of Engineering & Technology, Khulna, Bangladesh in October 2021.

BOARD OF EXAMINERS

1. _____ Chairman
   Dr. Md. Mahbub Hasan (Supervisor)
   Professor
   Department of Electrical and Electronic Engineering
   Khulna University of Engineering & Technology

2. _____ Member
   Head of the Department
   Department of Electrical and Electronic Engineering
   Khulna University of Engineering & Technology

3. _____ Member
   Dr. Md. Abdur Rafiq
   Professor
   Department of Electrical and Electronic Engineering
   Khulna University of Engineering & Technology

4. _____ Member
   Dr. Md. Shahjahan
   Professor
   Department of Electrical and Electronic Engineering
   Khulna University of Engineering & Technology

5. _____ Member
   Dr. Mohammed Imamul Hassan Bhuiyan (External)
   Professor
   Department of Electrical and Electronic Engineering
   Bangladesh University of Engineering and Technology

Acknowledgments

First and foremost, I am remarkably grateful to my creator Almighty Allah for giving me the strength to complete this thesis; it would not have been possible without His blessings.

I am immensely thankful to my respected supervisor, Professor Dr. Md. Mahbub Hasan, for his guidance and supervision that leads to the successful completion of this thesis. His timely direction, proper actions, and friendly attitudes made me complete all the thesis stages. I genuinely express my gratitude to him. In my life, I will not forget how hardship and difficulties he accomplished for timely arranging the final oral examination of this thesis. The respect and appreciation I have for him cannot be expressed in words. I feel fortunate to have such a friendly personality as my supervisor.

I cannot deny the motivation and direction on my career from another favorite teacher, Dr. Md. Arafat Hossain (Associate Professor, Department of EEE, KUET). I still remember his words and guidance when I was confused about continuing this M.Sc. Eng. program in KUET. Thank you so much for being there.

Furthermore, I am thankful to the respected board of examiners for their constructive feedback and suggestions on my thesis. On the final oral exam day, the presence of the respected examiner, Dr. Md. Abdur Rafiq (Professor, Department of EEE, KUET), even when he was undergoing heartbreaking grief, was genuinely inspirational and encouraging. I humbly express my gratitude to all of my teachers who helped and galvanized me in this journey.

Of course, my loving wife Tania and both of my parents hold a special recognition for their constant supports and encouragement. Your prayers always stay with me, which I believe have had a significant influence on this journey.

Indeed, this journey was challenging, but the support and motivation I received have undoubtedly made this possible.

October 2021 ***Md. Rakibul Hasan***

Abstract

Speech-related research has a wide range of applications. Although speech recognition has achieved significant success using integrated and efficient models, still some series of challenges remain as the linguistic-acoustic patterns are perturbed by speakers' individual articulation gestures and environmental noises. Temporally overlapping linguistic-acoustic features (i.e., formant trajectories) are found for the vocal tract dynamics in word pronunciation, whereas quasi-stationary and non-overlapped features are obtained for vowels. This thesis presents a comprehensive study on how Deep Neural Network (DNN)-based classifiers have marginalized variability due to speaker gestures and environmental noises, and how they improved classification performances focusing only on linguistic-acoustic patterns. Here, vocal tract resonance-based acoustic features, formant trajectories, are considered as the linguistic-acoustic features to investigate vocal tract dynamics. Vocal tract-induced variabilities are evaluated for both vowels and words by the coefficient of variations of the acoustical feature, and it is justified that the words have more variations than the vowels. Furthermore, an ANOVA (Analysis Of Variance) test has been performed on formant frequency-related features of a vowel and a word sample. Then, the statistical significance of all 14 formant frequency-related features is determined through Tukey's HSD (Honestly Significant Difference) test. This study also finds out the optimum number of Mel-Frequency Cepstral Coefficient (MFCC) features. Many speech-related researches employ MFCCs as acoustic features. However, finding the optimum number of MFCCs is an active research question. A 4-fold cross-validation approach is used in a DNN with *Adam* optimizer to compute performances in five different performance metrics, namely *confusion matrix, classification accuracy, Area Under Curve of Receiver Operating Characteristic (AUC-ROC), F_1 score,* and *Cohen's Kappa (κ)*. The same classification is performed by varying the input features and hidden layers of a general DNN architecture. Accordingly, the contributions of those individual feature sets and hidden layers are also identified. Experiments did not find any considerable contribution of formant transitions and dispersions in speech classification, and five hidden layers were optimum network configuration. In all different cases, this study has justified the hypothesis—word classification falls behind vowel classification due to variability. Furthermore, all performance metrics gave the best score for 24/25 MFCCs; hence this thesis suggests that the optimum number of MFCCs should be 25, although many ex-

isting studies use only 13 MFCCs. Also, it verifies that increasing the number of MFCCs yields better classification metrics with a lower computational burden than the increment of hidden layers. Using formant frequency, it has achieved as high as 89% classification accuracy and 99% AUC for vowels. For words, these scores were 64% and 91%, respectively. These scores are achieved with five hidden layer configurations having only 28,263 trainable parameters with five formant frequency features only. In the MFCC-based speech classification, the optimum number of MFCCs obtained from this study returns classification accuracies of 99% and 91% for vowel and word classification, respectively, where the vowel classification score outperforms state-of-the-art results. Such a good performance proves the efficacy of the proposed method.

Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ANOVA	Analysis Of Variance
AUC-ROC	Area Under Curve of Receiver Operating Characteristic
BERT	Bidirectional Encoder Representations from Transformers
COV	Coefficient Of Variation
CV	Consonant-Vowel
DCT	Discrete Cosine Transform
DF	Degrees of Freedom
DFT	Discrete Fourier Transform
DIVA	Directions Into Velocities of Articulators
DNN	Deep Neural Network
FACTS	Feedback Aware Control of Tasks in Speech
5F	Five Formant Frequency
5T	Five Formant Transitions
FN	False Negative
FP	False Positive
GMM	Gaussian Mixture Model
HL_1	One Hidden Layer
HL_2	Two Hidden Layers
HL_3	Three Hidden Layers
HL_4	Four Hidden Layers
HL_5	Five Hidden Layers
HL_6	Six Hidden Layers
HMM	Hidden Markov Model
HSD	Honestly Significant Difference
HTK	Hidden Markov Model Toolkit
LDA	Linear Discriminant Analysis
LPC	Linear Predictive Coding
LPCC	Linear Prediction Cepstral Coefficient
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficient

MS	Mean Squares
NLP	Natural Language Processing
PLP	Perceptual Linear Predictive
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SFC	State Feedback Control
SS	Sum of Squares
SVM	Support Vector Machine
TD	Task Dynamics
3F	First Three Formant Frequency (F1, F2, F3)
TN	True Negative
TP	True Positive
VCV	Vowel-Consonant-Vowel

CHAPTER I

Introduction

1.1 Background

Speech processing-based systems such as Microsoft Cortana, Google Assistant, and Amazon Alexa have vastly simplified our modern life. These command-based services are a helping hand not only for ordinary people but also for physically-challenged and old-age people. In general, speech recognition devices and systems have made several automations in our lives—home automation, automation in smartphones, laptops, and vehicles. Apart from these devices and systems, speech processing is making a significant impact in healthcare, such as diagnosis of several neurological diseases. Other applications include wearable for deaf persons [1], keyword spotting [2], accent classification [3], and language identification [4]. Accordingly, research employing speech processing is booming with a particular interest in speech recognition, classification, and generation.

Humans have God-gifted potential to recognize spontaneous or natural spoken languages with the highest accuracy. It is apparent that implementing such capability in machines will not be straightforward. Accordingly, with the help of advanced models like Deep Neural Network (DNN) (with a large number of data to train its parameters), less natural read-like speech recognition technologies have experienced compelling achievements in recent years. However, similar accomplishments on spontaneous speech recognition have not yet been achieved due to the contamination of linguistic-acoustic variables by non-linguistic spontaneous variables [5], [6]. The origin of such non-linguistic variables is related to speakers' age, gender, emotion, region, culture, and of course, vocal tract length. It is well-known that speech production is the overlap of comparatively stable vowel and dynamic consonantal articulation-related gestures of the vocal tract [7], [8]. To infuse information in speech, the

change of the vocal tract cavity with respect to time is called vocal tract dynamics. These vocal tract dynamics can be modeled by the time-varying filtering such as formant frequency, Mel-Frequency Cepstral Coefficient (MFCC), and Linear Predictive Coding (LPC).

Although vocal tract dynamics plays a significant role in speech classification, research reports, especially investigations on performance deviation due to vocal tract dynamics, are rarely found. This thesis presents a comprehensive study on the vocal tract dynamics-based classification performance deviation based on several feedforward DNN model configurations and input feature (formant frequency) combinations. Thus, it also reports the relative importance of formant frequency and its derived features (i.e., which features contribute more to speech recognition). The DNN-based classification model has accomplished the relegation of acoustic perturbation through proper tuning of its hyperparameters. Five classification metrics are utilized as a guiding tool to attain optimum DNN for classifying dynamic and quasi-stationary speech tokens (i.e., words and vowels).

In case of speech recognition, selection of appropriate feature is an important task. The performance of recognition depends heavily on the feature extraction phase because some important characteristics might be left out if it is not chosen correctly. Selection of the number of coefficients in any appropriate feature is as essential as the selection of classifiers, and in fact, classification performance employing MFCC is undoubtedly dependent on the optimum number of coefficients [9]. This thesis finds out the optimum number of MFCCs to extract the best possible score. Classification performance can be improved by incrementing the number of MFCCs and the number of hidden layers. More MFCC features indicate more acoustic information from speech. Therefore, when more MFCCs are used, DNN's computational burden increases for classifying information associated with additional MFCCs. Similarly, increasing the number of hidden layers also raises the computational burden of DNN. Therefore, a question arises: which one (MFCC increase vs. hidden layer increase) can produce a better result with a minimum computational burden? To the best of author's knowledge, very few works [9], [10] address similar questions and related issues. This thesis reports on this issue based on the Bengali language. Beside this, a more improved DNN-based classification model is developed to pull out the best possible classification score utilizing the optimum number of MFCC. Comparison with other relevant studies reveals the competitiveness of our classification scores. Since vowels and consonants are the fundamental building blocks of

any language, proper detection of vowels and consonants holds primary importance for any speech recognition system of that language. Moreover, a combination of vowels and consonants produces speech sounds that hold necessary information to communicate. Thus, the outcome of this research will help increase the performance of MFCC dependent systems.

This thesis particularly employs the Bengali language to show the speech-token classification. Bengali is the 6th most spoken language globally, and it has 267.7 million total users worldwide, including 228.7 million native speakers [11]. Despite that, extensive research in this language has not yet been done to such a level to efficiently use it in real communication devices as a speech-to-text recognizer.

1.2 Literature Review

1.2.1 Speech-Token Classification

Many researchers have been working on speech-token classifications. Tripathi *et al.* [12] performed classification of speech mode (read and conversational) for four Indian languages, including Bengali, by employing vocal tract feature in a single-layer feedforward neural network model. Furthermore, they employed the vocal tract features in multi-layer DNN-based phone recognition. Yang *et al.* [13] reported classification of ten English command words using three types of models, including feedforward DNN and CNN (Convolutional Neural Network). However, they did not explain classification performance. Similarly, Dawodi *et al.* [14] classified twenty Dari speech tokens employing CNN without any vocal tract dynamics-related study. Particularly for the Bengali language, Sharmin *et al.* [15] showed the classification of ten Bengali spoken digits using CNN. Das *et al.* [16] developed Bengali speech corpus for speaker independent continuous recognition, but they did not focus extensively on the recognition. Mandal *et al.* [17] demonstrated a Sphinx3-based Bengali speech recognition system with the main focus to help visually-impaired people. Islam *et al.* [18] performed isolated Bengali speech recognition based on CNN and RNN (Recurrent Neural Network). On top of these, Badhon *et al.* [19] remarked some state-of-the-art research in Bengali Speech Recognition up to the year of 2019, where recognition of Bengali vowels and consonants [20], detection of ten Bengali isolated words [21], recognition of Bengali

phonemes [22] are mentioned. Their study shows that the combination of the Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) is the most used classification model.

1.2.2 Classification Model

In previous decades, most speech recognition researchers utilized statistical HMMs [23] to consider the temporal variability of speech and GMMs to map the HMM states to the acoustic input features [24]. Although GMM-HMM models have several advantages, neural networks (DNNs) are continuously replacing their places because these DNNs can resolve some significant shortcomings of GMM-HMM models. In particular, HMM assumes the speech features as statistically independent, so it does not consider the possible correlation among individual features. Additionally, HMMs strongly depend on the arbitrary assumption of probability density function associated with states [25]. GMMs have some disadvantages as well, including statistical inefficiency for nonlinear data space [24]. On the other hand, DNNs have become very successful for modeling nonlinear data [24], and it is heavily used in pattern recognition tasks [26]. Several studies prove that DNNs with many hidden layers outperform GMM-based models by a large margin in various speech recognition benchmarks [24].

1.2.3 Acoustic Features

Different researchers have employed different sets of features for different speech-research purposes. Shanthi and Lingam [27] described state-of-the-art feature extraction techniques in speech research. Their research revealed that MFCC is the frequently used feature calculated from the short-term energy spectrum expressed on a Mel-frequency scale. It contains speakers' information and is commonly employed as a standalone feature in speech recognition tasks [13]–[15]. Linear Discriminant Analysis (LDA) is another technique that maximizes the between-class variation than the within-class variation in a dataset. Fusion MFCC, which is the combination of MFCC and LDA technique, was practiced by Gaikwad *et al.* [28]. LPC analysis is another technique that approximates the speech sample as a linear combination of past speech samples. Perceptual Linear Predictive (PLP) analysis extends the LPC technique, which is focused on cross-speaker isolated word recognition. Formant frequency is another important feature that is referred to as the peaks of the acoustic spectrum [29].

Formant Frequency

Formant frequency holds substantial acoustic information, and that is why many researchers have been utilizing this to classify vowels and vowel-consonant-vowel (VCV) sequences. Sounds come from our mouths due to resonance in our vocal tract, and formant frequencies refer to those resonant frequency peaks. The laryngeal energy activates the resonance, and these have a frequency of below 5 kHz in most cases. That is why the usual discussion of formant frequency lies up to the first five formant frequencies, and particularly, the vowels are often characterized by the first three formants [30]. In a study on the place of articulation, Story and Bunton [31] adopted the first three formant frequencies and their transitions, where they revealed the contribution of these transitions to the overall changes in the vocal tract shape at the time of speech production. They also provided direction that the time derivative of these transitions, which is basically the second derivative, is a possible means to estimate the contributions of both vowels and consonants to the variation of vocal tract dynamics. Kent and Read [30] emphasized formant transitions as an essential acoustic cue for speech perception, where second and third formant transitions are related to the place of production of a particular speech. Stephens and Holt [32] illustrated VCV and CV (consonant-vowel) utterances by morphing between natural speech tokens, and then the ability to classify those utterances was verified by human participants.

Along with formant frequencies, formant dispersion, i.e., the difference between two formants, is also used in several studies. Among these, López *et al.* [33] used 12-dimensional feature vectors, and Hasan *et al.* [34] used 9-dimensional feature vectors. Both of these research groups employed the first five formants and four dispersion between two different pairs of formants each time. Yusof *et al.* [35] used the first three formant frequencies and distances between each of them to classify Malaysian vowels, and they reported improvement for most of the vowels' classification when corresponding formant differences are incorporated. Some other studies also reported the classification of vowels using multiple combinations of the three lowest formants and computed differences among them [36], [37]. Yan and Vaseghi [38] used the first four formant frequencies to classify British, American, and Australian accents, and according to their findings, formant frequency plays a significant role in classifying accents.

Mel-Frequency Cepstral Coefficient (MFCC)

Rao and Manjunath [39] specified some standard features used in phone-based speech recognition systems. In general, the standard spectral features, including LPCC (Linear Prediction Cepstral Coefficient) and MFCC representing the gross shape of the vocal tract, are the most widely used. A recent survey paper [40] have reported 24 contemporary works on the Bengali language, out of which 15 works have used MFCC as input features. In another survey paper on Arabic speech recognition, Al-Anzi and AbuZeina [41] have reported 13 studies on isolated words speech recognition, out of which 12 studies have used MFCC. Furthermore, Silva *et al.* [10] have stated that the best recognition performance can be achieved by utilizing MFCC. Therefore, MFCC is the feature of significant interest in most speech-related researches.

The broad demand for MFCC can be proven from several studies in the Bengali language. Sharmin *et al.* [15] classified ten spoken Bengali digits by using MFCC features. In a Bengali speech corpus development, Das *et al.* [16] performed phoneme recognition using 13 base MFCCs and their first and second-order derivatives. Mandal *et al.* [17] also used the same features to develop a Sphinx3-based Bengali speech recognition system with the main focus to help visually-impaired people. Islam *et al.* [18] performed isolated Bengali speech recognition employing MFCC features. Furthermore, Sumon *et al.* [21] experimented classification of ten Bengali short speech words based on MFCC features and raw audio files in two different classification models. They found that MFCC-based classification outperforms raw audio-based classification. Syfullah *et al.* [20] performed recognition of Bengali speech characters (vowels and consonants) using MFCC features. On top of these, Badhon *et al.* [19] remarked some state-of-the-art studies in Bengali speech recognition up to the year of 2019. Their study reveals that MFCC and linear prediction coefficients are the most adopted feature extraction techniques.

Apart from the Bengali language, MFCC has been used in other language researches as well. Most of them utilized 13 MFCC features in a DNN classifier, such as recognition of speakers [42], English phoneme [43], emotion from English audio [44], five Malayalam vowel phonemes [45], twenty Dari speech tokens [14], three Arabic words [46], and ten English command words [13]. Some studies also preferred CNN models such as isolated English

word recognition by Soliman *et al.* [47]. Cen *et al.* [48] developed a real-time speech emotion recognition system from continuous speech utilizing three cepstral features—PLP cepstral coefficients, MFCC, and LPCC. Furthermore, Salau *et al.* [3] performed accent classification of three Nigerian languages using MFCC features. From their comparison with similar studies, it is evident that MFCC is the most accepted feature.

1.3 Importance of Vocal Tract Dynamics on DNN-Based Speech Classification

Where does the vocal tract dynamics carry importance? According to the prominent theory of coarticulation, the acoustic realization of a phonological segment depends on the context where it is being used [49]. This articulatory and acoustic context-sensitivity is associated with the dynamic interaction of the vocal tract, which essentially induces variation in acoustic features such as formant frequency and MFCC [50]. These dynamic articulatory activities are essential for conveying information to other humans and machines through speech tokens. In addition to the information transmission, these articulation dynamics provide a description of neurological diseases such as Parkinson's disease [51], [52], Dementia [53], and Alzheimer's disease [54]. These diseases cause vocal tract variation, which results in acoustical feature variation. Therefore, proper estimation of these vocal tract variations would help diagnose these diseases. Brabenec *et al.* [55] utilized 14 combinations of speech task and acoustic features for diagnosing Parkinson's disease. Hemmerling and Wojcik-Pedziwiatr [51] performed prediction and estimation of the same disease based on English vowel sounds. Additionally, Maor *et al.* [56] demonstrated an independent relationship of coronary artery disease with voice signal characteristics. Apart from these, the mechanics of vocal tract tube will guide the doctors in treatment of several other diseases like obstructive sleep apnea syndrome, dysphagia [57], dysphonia [58], cleft palate, and oral cancer [59]–[61]. Therefore, proper analysis of articulatory feature variations is necessary for automatic speech recognition and several biomedical applications [62]. Additionally, recognition of speech tokens has other applications, including emotion recognition [44], voice assistance in the smartphone, laptop, and vehicles, home automation [63], as well as assisting physically-challenged and old-age people [64], [65]. Thus, proper classification of speech tokens and the study of vocal tract dynamics have significant impact.

Performance evaluation of speech recognizers with respect to variability is a tremendously impactful task. Recognition of spontaneous and conversational speech requires consideration of internal structural information or generative mechanisms rather than only surface-level information. If only surface-level information is considered, theoretically, infinite information is required to completely cover the overwhelming variability [5]. That is why the variability induced by the vocal tract needs to be considered during speech recognizer design [66]. Vocal tract-induced formant variabilities are adapted for speech classification by the lifelong learning of humans. Based on the classification performance, Directions Into Velocities of Articulators (DIVA) [67], [68], Task Dynamics (TD) [69], [70], State Feedback Control (SFC) [71], and Feedback Aware Control of Tasks in Speech (FACTS) models have been formulated for computational speech-motor movement. These models are prominent in hearing impairment, stuttering, and phonatory learning. A well-trained DNN as a speech token classifier considering perturbed formant can be integrated effectively in the acoustics to sound mapping tool in the auditory feedback module of these models.

1.4 Scopes and Objectives

This study presents a step-by-step formation of Bengali speech classification model based on DNN. The hypothesis of this research is that there should have a noticeable impact on classification performance due to the variability induced by vocal tract dynamics. This research analyzes the reason behind such impact with a particular interest in the vocal tract dynamics. The specific objectives of the research are to—

i. construct an optimized DNN-based speech-token classification model,

ii. select and analyze appropriate speech features to use in DNN model,

iii. estimate vocal tract dynamics on speech classification,

iv. identify the importance of different speech features, and

v. analyze the performance of classification considering vocal tract dynamics.

1.5 Layout of the Thesis

This thesis is organized as follows. The following chapter (Chapter II) explains the Datasets, feature extraction process (both formant frequency and MFCC), and DNN-based classification model with hyperparameters and evaluation metrics. With the features and classification model, two separate types of experiments are reported in Chapter III and Chapter IV, respectively.

Chapter III starts with an overall workflow to investigate the effect of vocal tract dynamics on speech recognition followed by estimation of vocal tract dynamics. After that, several experiments by varying formant frequency-based input feature vectors and the number of hidden layers and hidden neurons are demonstrated. It presents the results of formant frequency-based speech classification with particular emphasis on observed classification performance deviation. Here, the performance comparison between vowel and word classification is analyzed with vocal tract dynamics.

Chapter IV also starts with an overall workflow to find out the optimum number of MFCC. This chapter can be broadly classified into four sections. Firstly, several experiments are reported to find the optimum number of MFCC. Secondly, the suitability between MFCC increase and hidden layer increase are compared. Thirdly, the DNN configuration is varied to find out the best possible score with the optimum number of MFCCs, and finally, similar existing works are compared based on both the performance score and the number of MFCCs used in those studies.

Lastly, all findings and applications are summarized in Chapter V with concluding remarks and some recommendations on future research directions.

CHAPTER II

Methodology

This chapter describes the Datasets, feature extraction, and DNN classifier involved in this work. Using these core methodologies, two separate types of experiments have been done which are reported with necessary flow diagrams (Figure 3.1, Figure 4.1) to show the key steps in corresponding chapters (Chapter III, Chapter IV).

2.1 Description of the Datasets

For the evaluation of the effect of vocal tract dynamics properties on the DNN-based classification and finding out optimum number of speech features, seven Bengali vowel-sounds (/অ/[/ɔ/], /আ/[/a/], /ই/[/i/], /উ/[/u/], /ঋ/[/ri/], /এ/[/e/], and /ঐ/[/oi/]) and seven word-sounds (বোতল, বন, কপি, দোকান, শেষ, সঠিক, and উপরে) were selected. These data were collected from 20 Bangladeshi speakers (age 20–26 years), who use Bengali as their first language. Guiding the speakers to pronounce the speeches in two different accents, 40 sound signals were captured corresponding to each of these seven vowels and words. The sounds were recorded on a 'Xiaomi Redmi 3' smartphone, and later they were processed in version 2.2.2 of the Audacity software [72]. The whole datasets used in this study are publicly available for research purpose [73]. It has 40 utterances in each of the seven classes for both vowels and words.

The volume of the datasets used in this work is comparable to other similar researches on speech classification, such as [20], [74], and [75]. Particularly for Bengali voice recognition, Syfullah *et al.* [20] used only 20 different sample inputs for each of their Bengali characters to classify. In a neural network-based word classification, Selvan and Rajesh [74] used 42 samples for each of their words. On the contrary, this thesis has utilized 40 different samples for each Bengali characters and words used in this study. Furthermore, Baquirin and Fer-

nandez [75] showed that 110 sound clips are reasonable for this type of works, whereas this thesis has used 280 sound clips for both vowel and word classification.

2.1.1 Why We Used These Specific Datasets

The Bengali alphabet has 11 vowels. Some vowels are not included in this study since almost all of those excluded vowels cannot be separated in spoken Bengali language. For example, while pronouncing words consisting of vowels, ই cannot be separated from ঈ, and উ cannot be separated from ঊ. They actually have a very similar pronunciation while used in words of spoken language. That is why one vowel from each pair (e.g. ই from ই, ঈ) is selected in the vowel dataset.

To be consistent, since seven vowels were selected, we searched for seven words (as our primary target is to compare vowel and word classification). Those particular seven words were selected since they tend to vary when pronounced by different speakers. Those selected words have diverse pronunciations, which means their pronunciation is not quite fixed across different speakers. A speaker might say the word বোতল differently than what another speaker might pronounce. The same thing is true for all of those seven words. Since this thesis aims to study variations, we choose those specific words.

2.2 Feature Extraction

Both formant frequency and MFCC have been utilized as speech features in the classification model. Furthermore, the optimum number of speech features for a competent classification performance are also experimented with these features. Raw speech to corresponding feature extraction is explained in the following subsections.

2.2.1 Formant Frequency

Samples at every 6 ms interval are taken to extract the five formant frequencies in a window length of 25 ms using PRAAT script [76]. Additionally, four specific formant dispersions, which are the differences between two formant frequencies, are calculated for all vowels and

words according to Equation 2.1.

$$\begin{aligned} F51 &= F5 - F1 \\ F43 &= F4 - F3 \\ F53 &= F5 - F3 \\ F54 &= F5 - F4 \end{aligned} \tag{2.1}$$

On top of these, five formant transitions across the time axis are calculated by taking the second derivative with backward difference approximation given in Equation 2.2.

$$F_i'' = |F_i - 2F_{i-1} + F_{i-2}| \tag{2.2}$$

where, i denotes a counter that goes from 1 to the maximum number of formant values for each of the five fundamental formants in each class. By applying the above equation to all five fundamental formants, five more features related to formant frequency are extracted.

2.2.2 Mel-Frequency Cepstral Coefficient (MFCC)

This thesis has utilized version 0.7.2 of the librosa package [77] in the Python (version 3.x) programming language to extract the MFCC features. MFCC feature extraction process involves applying Discrete Fourier Transform (DFT) on a signal window, taking the logarithm, and then expressing on a Mel scale, followed by a Discrete Cosine Transform (DCT). Then the DCT components refer to the Mel-frequency cepstral coefficients or MFCCs. Rao and Manjunath [39] presents a wonderful explanation of the process from where the key steps are briefly explained as follows.

Often a filtering is performed to emphasize on higher frequency components. Transfer function of such a widely used pre-emphasis filter is given in Equation 2.3.

$$H(z) = 1 - bz^{-1} \tag{2.3}$$

where, b, having a typical value between 0.4 to 1, represents the slope of the filter.

In the next step, the speech signal is segmented typically on a 20 ms window advanced every 10 ms, and for this windowing, Hanning or Hamming window are generally utilized. Then, DFT is applied to these windowed frames, which results in magnitude spectrums according to Equation 2.4.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}} \; ; \qquad 0 \leq k \leq N-1 \tag{2.4}$$

where, N denotes the number of points in DFT.

These Fourier-transformed signal is passed through a Mel-filter bank, the fundamental band-pass filter behind MFCC computation. Mel spectrum is achieved from physical frequency according to the approximation given in Equation 2.5.

$$f_{Mel} = 2595 \log_{10} \left(1 + \frac{f}{700}\right) \tag{2.5}$$

where, f_{Mel} is the converted Mel-spectrum from physical frequency f.

Accordingly, the multiplication of magnitude spectrum $X(k)$ by triangular Mel weighting filters results in the Mel-spectrum of $X(k)$, which is shown in Equation 2.6.

$$s(m) = \sum_{k=0}^{N-1} \left[|X(k)|^2 H_m(k) \right] \; ; \qquad 0 \leq m \leq M-1 \tag{2.6}$$

where, the total number of weighting Mel-filters is denoted by M. The k^{th} spectrum bin is weighted by $H_m(k)$, which results in m^{th} output Mel-frequency band. $H_m(k)$ can be expressed according to Equation 2.7.

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \leq k \leq f(m) \\ \frac{2(f(m+1)-k)}{f(m+1)-f(m)}, & f(m) < k \leq f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{2.7}$$

where, m takes a value between 0 to $M-1$.

Finally, DCT is applied to the Mel-frequency band which results in cepstral coefficients. Before DCT, the Mel-frequencies are generally converted to a logarithm scale as illustrated

in Equation 2.8.

$$c(n) = \sum_{m=0}^{M-1} \log_{10}\left(s(m)\right) \cos\left(\frac{\pi n(m-0.5)}{M}\right); \qquad n = 0, 1, 2, \ldots, C-1 \tag{2.8}$$

where, c refers to the ultimate Mel-frequency cepstral coefficients or MFCCs, and C denotes the number of coefficients.

2.3 Deep Neural Network-Based Classification Model

A feedforward DNN is utilized as the main classifier to classify vowels and words. Sometimes more advanced DNN architectures (such as CNN) are used for these types of classification tasks. Particularly for Natural Language Processing (NLP) tasks, RNNs (such as LSTM—Long Short-Term Memory) or more advanced big Transformer models (such as BERT—Bidirectional Encoder Representations from Transformers) are widely used. It is worth mentioning that although we are dealing with a natural language, here we are observing the effect of vocal tract dynamics on speech token classification rather than conventional NLP tasks like next-word prediction or text summarization. Therefore, feedforward DNN is more suitable than models like LSTM and BERT for this purpose. Someone might ask why did this thesis use such a simpler DNN model? First of all, we are more interested in studying the effect of vocal tract dynamics on vowel and word classification performance. A classification model like CNN would do the job, but the main objective of this research (effect of acoustic variation on classification) cannot be clearly estimated as the convolution operation of CNN relegates those effects. Secondly, a feedforward DNN is the building block of all these advanced models. In a typical CNN architecture, feedforward DNNs are placed after the initial convolution layers. Furthermore, more recently, researchers are focusing on these simpler DNNs (also called multilayer perceptrons) for various classification tasks. Touvron *et al.* [78] efficiently solved a complex classification task employing simpler DNNs (also called multilayer perceptrons) rather than a CNN. Moreover, a more simplistic and efficient model is preferable for deploying the model in low-computation-capable edge devices such as smartphones, Arduino, microcontrollers, and Raspberry Pi. Therefore, a basic feedforward DNN model is selected as the primary classifier to study vocal tract dynamics. A similar study on speech classification [12] also used such a basic model.

The following subsections explain the elements involved in the classification model.

2.3.1 Feature Standardization

Extracted formant frequency and MFCC features have different dimensions having values at different ranges. Feature standardization or Z-score normalization is a method to transform the data to zero-mean and unit-variance that essentially helps DNNs and other machine learning algorithms to converge faster [79]. Accordingly, all acoustical features are standardized before feeding in their respective models according to Equation 2.9.

$$x'_c = \frac{x_c - \overline{x_c}}{\sigma_c} \tag{2.9}$$

Here, the subscript c indicates that this normalization operation is performed to each column, where the columns represent different dimensions (five formants, five dispersions, and four transitions for formant frequency; MFCC numbers for MFCC feature). $\overline{x_c}$ and σ_c denotes the mean and the standard deviation of that column, and x'_c is the standard features having a mean of zero and variance of one. These normalized features are supplied in the input layer of the feedforward DNN-based classification model.

2.3.2 Model Configuration

The number of neurons in the input layer depends on the number of input features. The vowel and word classes act as the output layers for the vowel and the word classification models, respectively. In between the input and output layer, there could be one or many hidden layers. The number of hidden layers and the number of neurons in these hidden layers cannot be defined by any hard and fast rule or formula. It varies based on the application area and can be best approximated by multiple experiments. This thesis came up with five hidden layers as the optimum number through experiments. Nevertheless, classifications using one, two, three, four, and six hidden layers are also reported to justify choosing five hidden layers.

Experiments on Vocal Tract Dynamics

In the investigation into the effect of vocal tract dynamics, the number of input feature vectors used in classification are varied. Classification is performed employing five different sets of input features—only five input vectors (five formants), only three input vectors (first three formants), ten input vectors (five formants and five dispersions), nine input vectors (five formants and four transitions), and fourteen input vectors (all features—five formants, five dispersions, and four transitions). Thus, performances are compared among combinations of formant frequencies, formant dispersions, and formant transitions. The architecture of the feedforward DNN model having five hidden layers and five formants in input, particularly for vowel classification, is shown in Figure 2.1.

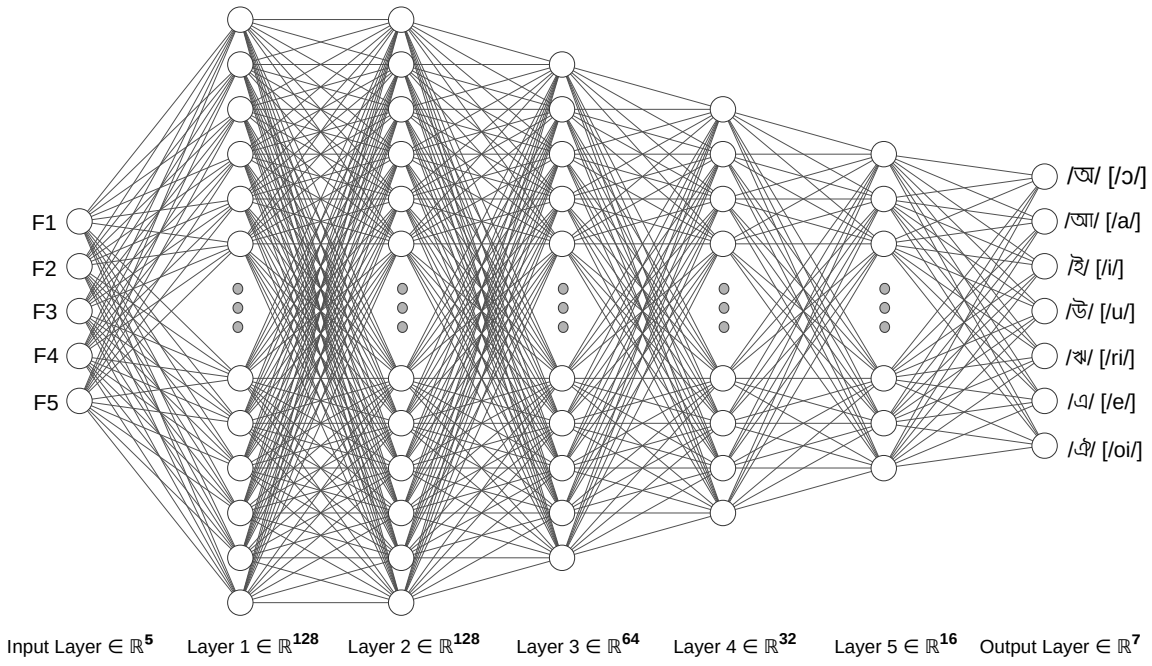


Figure 2.1: The architecture of the fully-connected feedforward DNN-based vowel classification model having five hidden layers. Input neurons represent the five formant frequency features, and output neurons represent the output vowel classes.

In the figure, five neurons in the input layer are involved as five formant frequencies (input feature vectors). When all 14 feature vectors (five formants, five dispersions, and four transitions) are employed, there would be 14 neurons in the input layer. Similarly, when it is a word classification model, the output classes would be words rather than vowels as shown in Figure 2.1.

Experiments on Optimum Number of MFCCs

The number of input MFCC features are varied to find out the optimum number of MFCC features. Two hidden layers of 32 and 16 neurons, respectively are utilized to find out the optimum number of MFCC, the architecture of which is presented in Figure 2.2 particularly for vowel classification. As the number of input features are varied during the optimum number of MFCC searching, the input neurons of the Figure 2.2 changed accordingly. Plus, for word classification, the output classes correspond to seven words. Furthermore, after finding the optimum number of MFCC features, the number of hidden layers and the number of neurons on them are varied to obtain the best possible score.

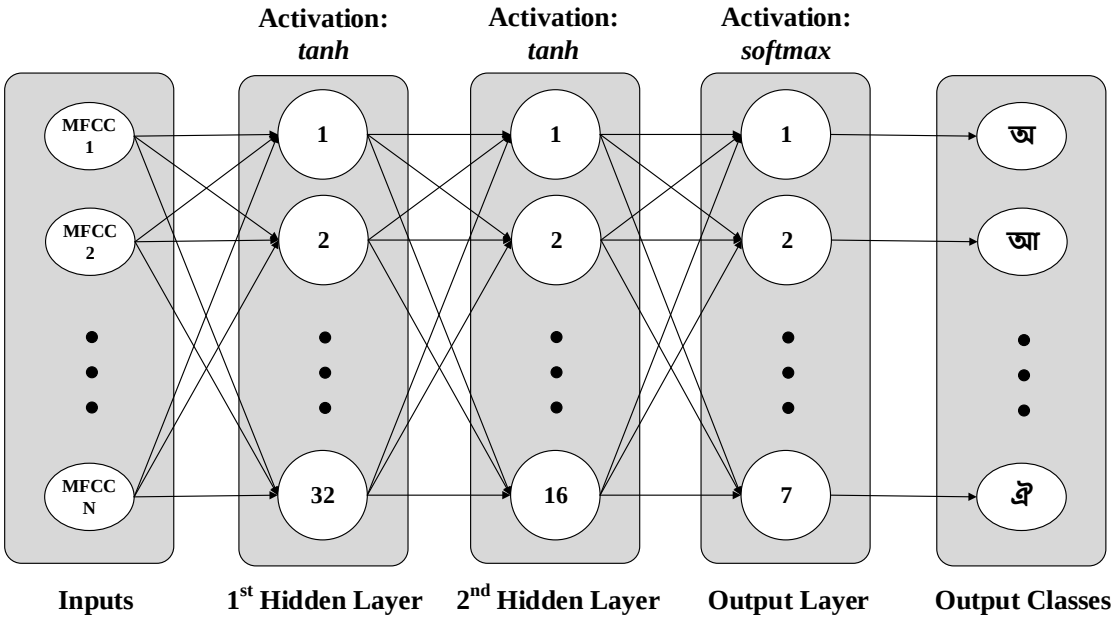


Figure 2.2: The architecture of a fully-connected DNN-based vowel classification model having two hidden layers. Input neurons represent the MFCC features, and output neurons represent the output vowel classes of the vowel classification model.

2.3.3 Weights, Biases, and Hyperparameters

Weights and biases are two fundamental trainable parameters of any DNN-based model. In any dense layer, the output of any neuron is calculated according to Equation 2.10.

$$\begin{aligned} v &= b + \sum_{i=1}^{n} x_i \cdot w_i \\ y &= F\left(v\right) \end{aligned} \tag{2.10}$$

where, $F =$ activation function, $b =$ bias, $w =$ weights, $x =$ input to neuron, $n =$ the number of inputs from the incoming layer, and v refers to the linear operation that is applied to each and every neurons in the model. After the linear operation, an activation function is applied to all neurons in a layer to adapt to the system's non-linearity and restrict the output values within a certain limit defined by that particular activation function type. In hidden layers, *tanh* activation function is chosen since it gave a better performance while experimenting with several activation functions such as *ReLU (Rectified Linear Unit)* and *Leaky ReLU*. The *tanh* activation limits the output from -1 to $+1$ as shown in Equation 2.11.

$$F\left(v\right) = \tanh\left(v\right) = \frac{e^{v} - e^{-v}}{e^{v} + e^{-v}} \tag{2.11}$$

For a multi-class classification model like what we are dealing with, *Softmax* activation function is an optimal choice in the output layer. It returns the probabilities of each class through which the target class is determined, having the highest probability close to one [80]. The mathematical equation is given in Equation 2.12.

$$\hat{y}_j = F\left(v_j\right) = \frac{e^{v_j}}{\sum_{j=1}^{K} e^{v_j}}; \quad j = 1, 2, 3, \cdots K \tag{2.12}$$

where, K is the total number of output classes, and $\hat{y}_j$ denotes the prediction for j^{th} class.

At the training phase, example input-output patterns are served to the model so that it can tune its trainable parameters to predict the output of the classification task. Presentation of input features to the model can be either in sequential mode or in batch mode. Sequential mode is also called stochastic mode, where individual training samples are passed from the

input node to the output node one by one, and then the trainable parameters are perturbed with respect to these individual sequences. On the contrary, batch mode training involves all samples to be passed at once, and accordingly, the parameters are also perturbed based on those training samples at once. For a large dataset, the sequential mode is an optimal choice as it needs less computational power [81]. In this thesis, batch mode training is chosen.

One pass of the full training set through the model is termed as one *epoch*. The whole dataset (both vowel and word) is divided into a training set of 80% and a validation set of 20%. In each *epoch*, all of these trainable parameters are slightly changed towards the best values by minimizing the loss function. This loss function is actually a measure of the difference between the actual output and predicted output. This study has used *categorical cross-entropy* loss function as defined by Equation 2.13.

$$\text{Loss} = -\sum_{j=1}^{K} y_j \cdot \log \hat{y}_j \tag{2.13}$$

where, y_j is the ground-truth or target value, and $\hat{y}_j$ denotes the prediction made by the model for j^{th} class. More deviation between these two will result in a higher loss score. The key technique behind optimizing a DNN is to use this score as a feedback signal to fine-tune the trainable parameters gradually to reduce this loss score. An optimizer makes this adjustment through which it implements the fundamental *Backpropagation* algorithm. The optimizer used in this work is *Adam* with a learning rate of 0.005. It is one of the heavily used optimization algorithms in the deep learning domain [82], [83].

The number of weights equals the number of connections between neurons of a layer and its preceding layer, whereas the number of bias parameters is equal to the number of neurons in that particular dense layer. Since this is a fully-connected network, all input neurons are fully connected to the first hidden layer, all neurons of the first hidden layers are fully connected to the second hidden layer, and so on. The total weight parameter (say, W_l) in a dense layer (say, l) can be found by multiplying the number of neurons (say, n_l) by the number of input to that particular layer (i.e., the number of neurons in the preceding layer, n_{l-1}). The total bias parameters (say, B_l) is equal to the number of neurons (n_l) of that particular layer l. Therefore, the total number of trainable parameters in a particular layer l can be derived as shown in Equation 2.14. Particularly for the configuration shown in Figure 2.1, the total

trainable parameters is 28,263.

$$\begin{aligned} W_l &= n_l \times n_{l-1} \\ B_l &= n_l \\ \text{Thus, the total trainable parameters} &= W_l + B_l \end{aligned} \tag{2.14}$$

2.3.4 Evaluation Metrics

Some specific metrics are required to compare the performance between vowel and word classification. To measure the performances of the classification, this thesis utilizes five different metrics—*Confusion matrix*, *classification accuracy*, *Area Under Curve of Receiver Operating Characteristic (AUC-ROC)*, *F_1 score*, and *Cohen's κ*. For all these different evaluation metrics mentioned above, a higher value implies better classification performance.

A typical confusion matrix is shown in Figure 2.3. It has four primary elements.

	actual positive	actual negative
predicted positive	TP	FP
predicted negative	FN	TN

Figure 2.3: Elements of a confusion matrix.

True Positive (TP) refers to actual positive examples predicted as positive as well, whereas False Positive (FP) is actually negative, but the classifier predicts as positive. Similarly, True Negative (TN) are actual negative predicted as negative as well, whereas False Negative (FN) are actual positive cases predicted as negative by the classifier. Several evaluation metrics can be defined from the confusion matrix [84], some of which are given in Equation 2.15.

$$\begin{aligned} \text{Classification accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\ \text{Precision or True Positive Rate} &= \frac{TP}{TP + FP} \\ \text{False Positive Rate} &= \frac{FP}{FP + TN} \\ \text{Recall} &= \frac{TP}{TP + FN} \end{aligned} \tag{2.15}$$

Classification accuracy measures the correctness of classification, which means it gives a ratio of correct predictions (both TP and TN) to all predictions the model made. Since the true label of the classes is known for both training and validation time, the algorithm compares this true label with the predicted label to provide this score. When the number of samples is different in different classes, the *accuracy* score cannot explain the whole scenario. In that case, *AUC-ROC* measures the degree of separability among different classes. A Receiver Operating Characteristic (ROC) curve is a plot of *True Positive Rate* versus *False Positive Rate*. As discussed previously, the *Area Under Curve of ROC (AUC-ROC)* is a means to evaluate a classification performance, especially when the sample distribution is not uniform.

Precision or *True Positive Rate* is the ratio of the number of correctly classified labels to all predictions the model picked as correct, including those not identified correctly, and *Recall* or *Sensitivity* is the ratio of the number of correctly classified labels to all labels that should have been classified correctly (i.e., all ground truths). In most cases, an inverse proportional relationship exists between *Precision* and *Recall*, and therefore, a harmonic mean of these two metrics gives a better estimate of the model's performance. This metric is known as F_1 *score* or *F-measure* which is calculated according to Equation 2.16.

$$\mathrm{F}_1 \text{ score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.16}$$

Thus, F_1 *score* combines two metrics—*Precision* and *Recall*, which are proportional to correct classification. Other than these derived metrics, *confusion matrix* itself is used as another graphical performance metric that facilitates a way of observing classification performance with respect to individual labels or classes.

Apart from these metrics, Cohen's κ is a statistical metric that measures the inter-rater agreement, and it tells us how much the classifier is performing other than a model that delivers just a random guess [85]. According to Landis and Koch [86], a value of less than zero tells that the model is giving a random guess; a value of 0.81 to 1.00 implies an almost perfect model; a value of 0.61 to 0.8 indicates a substantial-good model, and so on.

Different performance metrics have different pros and cons, and a model's performance should not be justified based on a single metric. That is why, all these five performance

metrics are considered. However, the value of these metrics varies due to different reasons, including random initialization of network parameters. Furthermore, it depends on which portions of data are used for the training phase and which portions are for the validation phase. Our experiments arbitrarily decided which sample data to be used in which phase so that the performance cannot be biased to sample data. Thus, the metrics' values changed for different runs. K-fold cross-validation is a well-known technique to combat this issue, in which the total dataset is split into K number of folds. In each run, $(K-1)$ folds are used to train the model, and the remaining single fold is used to validate the model [87]. Thus, K runs are required to traverse all the folds in total, and then the average performance of all of these folds is used as a cross-validated evaluation metric.

CHAPTER III

Experimental Results: Effect of Vocal Tract Dynamics

The following experiments utilize formant frequency as acoustical feature. A sketch of the steps involved in analyzing the effect of vocal tract dynamics and finding out the importance of different formant frequency features is depicted in Figure 3.1.

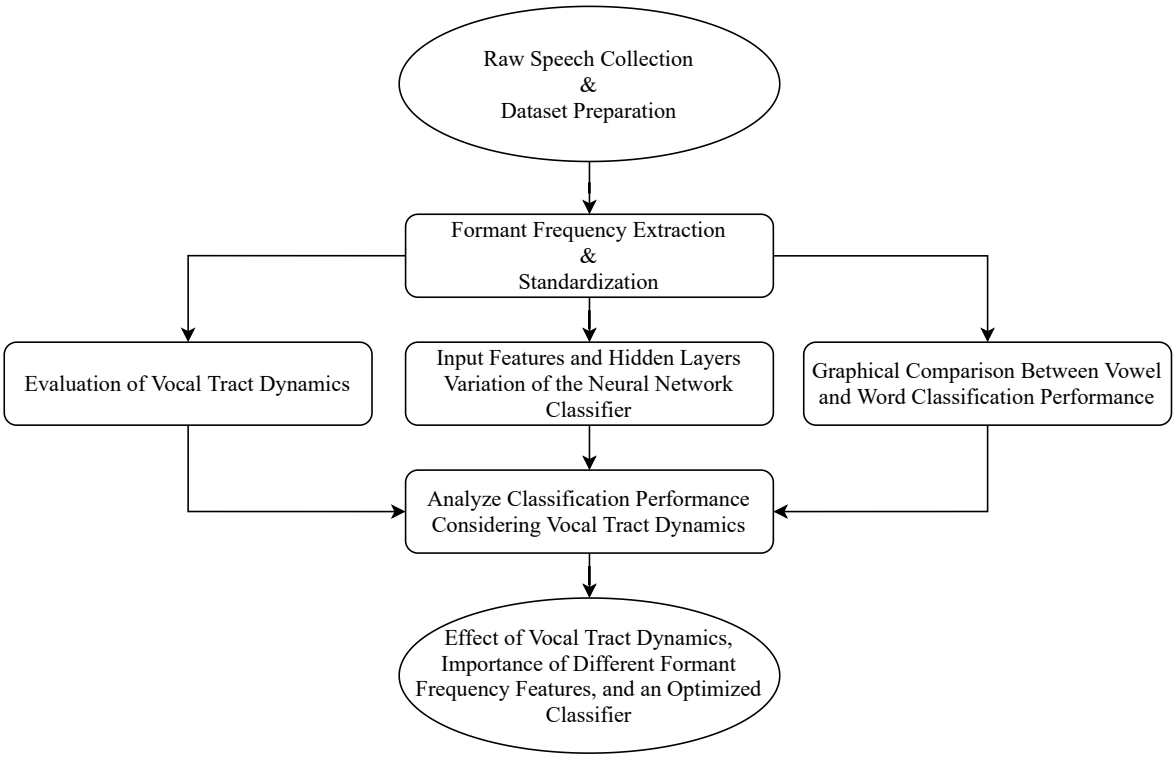
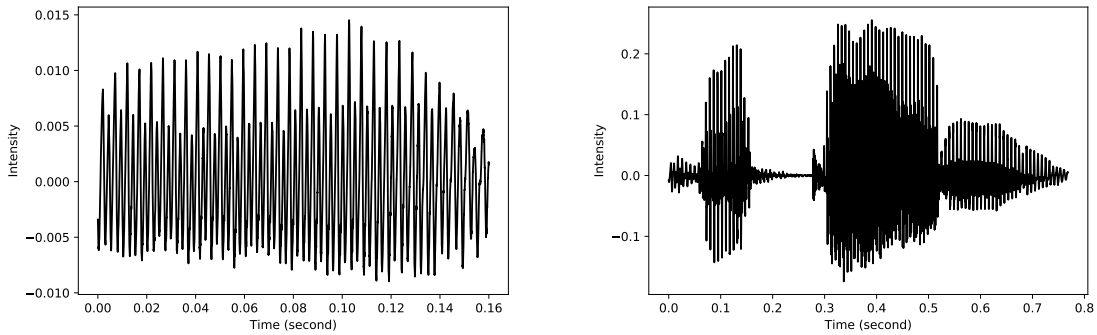


Figure 3.1: Steps involved in analyzing the effect of vocal tract dynamics and finding out the importance of different formant frequency features.

3.1 Evaluation of Vocal Tract Dynamics

To demonstrate the shape of vowel and word waves in the time domain, the waveforms of a sample vowel (/ঊ/[/u/]) and a sample word (দোকান) are shown in Figure 3.2.

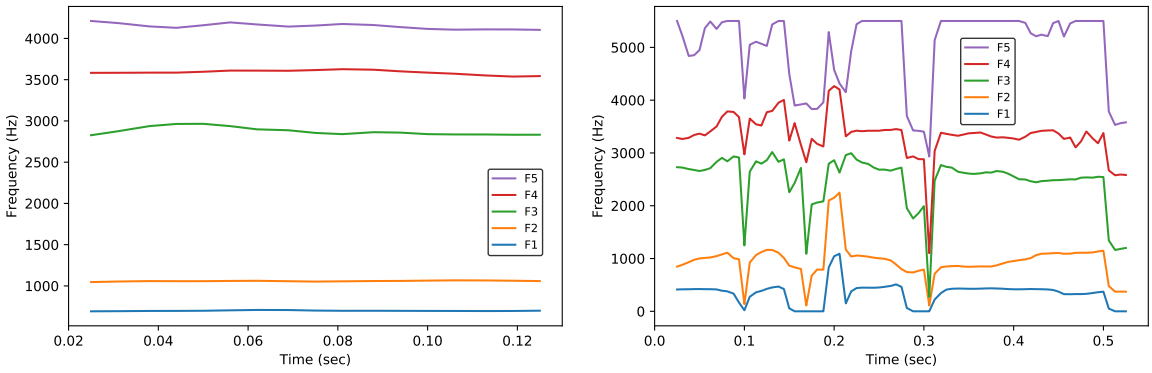


(a) উ vowel. (b) দোকান word.

Figure 3.2: Waveshapes of a vowel and a word uttered by a particular speaker. Word waveshape shows more discontinuity than that of the vowel.

Figure 3.2a and Figure 3.2b depict that the vowel waveform is relatively steady, but silence presents within the word waveform since the word consists of vowels and consonants. The glottal pulse source energizes the quasi-stationary vocal tract during vowel production, and the output becomes an approximately steady-state waveform. On the contrary, consonantal-constrictions are present in word production, which induces the transitional nature of the vocal tract, and thus the output becomes a discontinuous waveform shown in Figure 3.2b.

Formant trajectories are usually used to illustrate the state of the transfer function of vocal tract during speech production, and the dispersion of formant trajectories exhibits the numerical values of vocal tract dynamics [88]. For visual comparison, the formant trajectories of /অ/ [/ɔ/] vowel and বোতল word are shown in Figure 3.3a and Figure 3.3b respectively.



(a) অ vowel. (b) বোতল word.

Figure 3.3: Formant trajectories for a vowel and a word. There are more variations in the word formants as compared to vowel formants.

Between Figure 3.3a and Figure 3.3b, more dispersive nature is observed in Figure 3.3b due to

the presence of several consonantal constrictions during word production. These consonantal constrictions in the vocal tract induce dispersive behavior in vocal tract resonance as well as filtering properties. The Coefficient Of Variation (COV) of all formants is calculated for all vowels and words. To calculate this variation for a particular formant for a particular vowel or word, at first, mean (μ), standard deviation (σ), and COV are calculated for a single source (speaker) according to the Equation 3.1.

$$\begin{aligned} \mu_j &= \frac{\sum_{i=1}^{n} x_i}{n} \\ \sigma_j &= \frac{\sqrt{\sum_{i=1}^{n} (x_i - \mu_j)^2}}{n-1} \\ COV_j &= \frac{\sigma_j}{\mu_j} \end{aligned} \tag{3.1}$$

where, n = total number of formant values for the j^{th} source. Considering N_s number of sources for that particular vowel or word, the overall COV of that particular formant is calculated using Equation 3.2.

$$COV = \frac{\sum_{j=1}^{N_s} COV_j}{N_s} \tag{3.2}$$

The average COVs as calculated from Equation 3.2 using $N_s = 40$ sources are shown in Table 3.1 for both vowels and words. Here, we can easily observe that there is a sharp decrease in the magnitude of variation with increasing formant order up to F4. The variations between F4 and F5, especially for vowels, are less discernible. Between words and vowels, words revealed higher valued COVs. Also, the average coefficient of variation from Table 3.1 is shown in Figure 3.4 for visual comparison. It clearly illustrates that the words' formants have more variation than vowels. The first formants have the highest dispersive nature, and the variation decreases for lower-order formants.

3.2 Analysis of Variance and Formant Frequency Feature Selection

This section provides some statistical analysis on the 14 formant frequency-related features of a sample vowel (/অ/[/ɔ/]) and a sample word (বোতল). For both of these samples, all speakers' data are concatenated one after another. The analysis provides a direction on which set of features are statistically significant and which are not.

Table 3.1: Coefficient of variation (COV) for the vowel and word formant frequencies. Higher values of COV denote higher variation in acoustical features (formant frequency).

	Coefficient of Variation (COV)				
Speech	**F1**	**F2**	**F3**	**F4**	**F5**
অ	0.2959	0.1823	0.1064	0.0668	0.0603
আ	0.3443	0.0896	0.0853	0.0602	0.0599
ই	0.3288	0.2575	0.0670	0.0536	0.0669
উ	0.4756	0.2167	0.1558	0.0672	0.0719
ঋ	0.4230	0.3123	0.1077	0.0829	0.0874
এ	0.3340	0.2063	0.0659	0.0451	0.0647
ঐ	0.3915	0.4939	0.1312	0.0623	0.0701
বোতল	0.4623	0.3073	0.1633	0.1004	0.0869
বন	0.6066	0.3503	0.2087	0.1226	0.1050
কপি	0.7004	0.4955	0.1813	0.1027	0.0969
দোকান	0.6829	0.3615	0.1936	0.1266	0.0926
শেষ	0.9849	0.2661	0.1335	0.0847	0.0710
সঠিক	0.9099	0.3926	0.1916	0.1133	0.0940
উপরে	0.5334	0.4157	0.1753	0.1197	0.0932

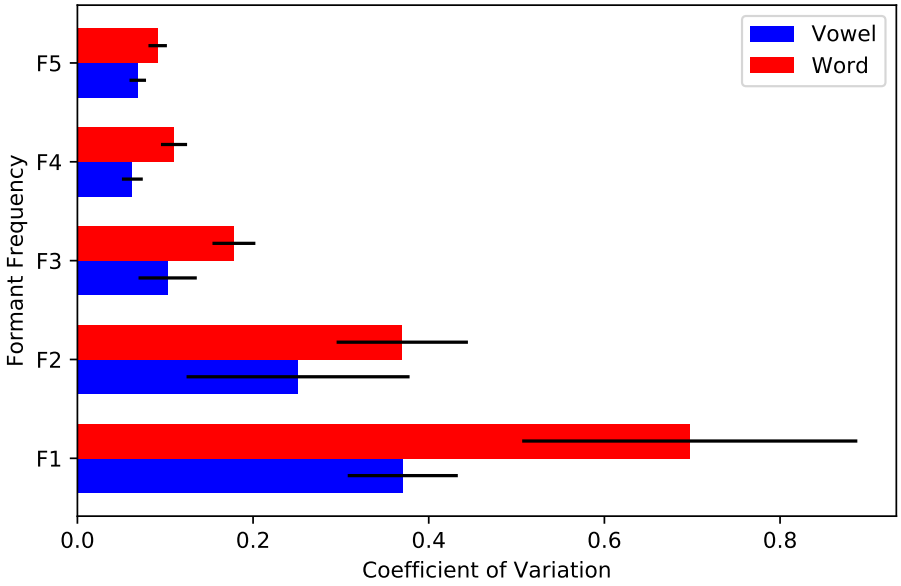


Figure 3.4: Average formant COV comparison between vowel and word (deduced from Table 3.1). Words have more variability than vowels.

3.2.1 Summary Statistics

Table 3.2 and Table 3.3 report total number of sample count, sum, average, and variance of each of the formant frequency-related features for the /অ/[/ɔ/] vowel বোতল word, respectively.

Table 3.2: Summary statistics of formant frequency-related features of /অ/[/ɔ/] vowel.

Feature	Count	Sum	Average	Variance
F1	1484	866250.98	583.7270755	54219.22565
F2	1484	1669570.88	1125.047763	154094.7943
F3	1484	3688565.93	2485.556557	156708.4292
F4	1484	5451213.7	3673.324596	202983.4683
F5	1484	6882419.08	4637.748706	218200.3812
F1″	1484	113339.84	76.37455526	35493.1726
F2″	1484	163873.29	110.4267453	63459.26143
F3″	1484	215601.16	145.2838005	144303.1826
F4″	1484	269863.58	181.8487736	124797.2352
F5″	1484	304167.03	204.9643059	144386.5312
F51	1484	6016168.1	4054.021631	185334.0068
F43	1484	1762647.77	1187.768039	137152.62
F53	1484	3193853.15	2152.19215	183468.9797
F54	1484	1431205.38	964.4241105	155252.2279

Table 3.3: Summary statistics of formant frequency-related features of বোতল word.

Groups	Count	Sum	Average	Variance
F1	3040	1148154.48	377.6823947	30773.40048
F2	3040	3116011.44	1025.003763	108160.1025
F3	3040	7669982.91	2523.020694	221057.7689
F4	3040	11131285.48	3661.607066	174338.2556
F5	3040	14436984.34	4749.008007	238352.6907
F1″	3040	231307.74	76.08807237	24563.10201
F2″	3040	495450.21	162.9770428	90345.63048
F3″	3040	719521.08	236.6845658	194674.2052
F4″	3040	665712.29	218.9843059	128230.998
F5″	3040	773647.86	254.4894276	146141.255
F51	3040	13288829.86	4371.325612	203910.0092
F43	3040	3461302.57	1138.586372	151192.0733
F53	3040	6767001.43	2225.987313	281385.1082
F54	3040	3305698.86	1087.400941	248551.0132

3.2.2 ANOVA Test

Table 3.4 presents one-way ANOVA (Aalysis Of Variance) test result on the selected vowel. This test has been performed having the significance level $\alpha = 0.05$. In the ANOVA test,

the null hypothesis (all features have the same mean, which further specifies that the data are not statistically significant) can be rejected if the computed p-value is less the α. Rejection of null hypothesis means overall the features are statistically significant.

Table 3.4: ANOVA test result on the formant frequency-related features of /অ/[/ɔ/] vowel.

Source of Variation	SS	DF	MS	F	p-value	F critical
Between Groups	49025796372	13	3771215106	26939.26411	0	1.720627056
Within Groups	2906462764	20762	139989.5368			

SS: Sum of Squares; DF: Degrees of Freedom; MS: Mean Squares

Here in Table 3.4, the computed p-value is less than α, which means that the features are statistically significant. Similar to the vowel sample, the same ANOVA test has been performed on the selected word sample which is illustrated on Table 3.5.

Table 3.5: ANOVA test result on the formant frequency-related features of বোতল word.

Source of Variation	SS	DF	MS	F	p-value	F critical
Between Groups	1.07459E+11	13	8266089469	51624.44196	0	1.720386039
Within Groups	6812452188	42546	160119.6866			

SS: Sum of Squares; DF: Degrees of Freedom; MS: Mean Squares

Here in Table 3.5 for the sample word, we can again observe that the p-value is less than α. Therefore, overall, the features are statistically significant.

3.2.3 Tukey's HSD Test and Feature Selection

ANOVA test provides overall results on whether the features are statistically significant or not. Furthermore, to know which particular features are statistically significant, a post-hoc test needs to be performed. Among several post-hoc tests, Tukey's range test or Tukey's HSD (Honestly Significant Difference) test compares all features pairwise and provides the significance level on those individual pairs. Since we have 14 formant frequency-related features, there are ${}^{14}C_2 = 91$ possible pairs for each of the vowels and words. Tukey's HSD test rejected the null hypothesis on 87 pairs among all these 91 pairs for the vowel sample (Table 3.6).

Table 3.6: Selected Tukey's HSD pairwise comparison result of vowel formant frequency where the null hypothesis is not rejected, which means these four pairs are not statistically significant.

Group 1	Group 2	Mean diff	p-adjusted	Lower	Upper	Reject?
F1″	F2″	34.0522	0.4271	-12.019	80.1233	FALSE
F2″	F3″	34.8571	0.3848	-11.2141	80.9282	FALSE
F3″	F4″	36.565	0.3017	-9.5062	82.6361	FALSE
F4″	F5″	23.1155	0.9	-22.9556	69.1867	FALSE

The remaining four pairs mentioned in Table 3.6 are not statistically significant (the null hypothesis is not rejected). It can be further verified since the adjusted p-value is not less than α for these four pairs. Except for the four pairs mentioned in the table, the null hypothesis is rejected for all other 87 pairs (they have the p-adjusted value of around 0.001), which means all those other pairs are statistically significant. Therefore, these four pairs, which are formant transitions features, are not statistically significant in the sample vowel.

Table 3.7 reports three selected results from Tukey's HSD test on the word sample's formant frequency-related features. Except for the first two pairs mentioned in the table, the null hypothesis is rejected for all other 89 pairs, which means all those other pairs are statistically significant considering $\alpha = 0.05$. The third pair in Table 3.7 reflects that if we had chosen $\alpha \leq 0.0357$, this pair would have also been considered insignificant.

Table 3.7: Selected Tukey's HSD pairwise comparison result of word formant frequency where the null hypothesis is not rejected for the first two pairs mentioned here, which means these two pairs are not statistically significant.

Group 1	Group 2	Mean diff	p-adjusted	Lower	Upper	Reject?
F3″	F4″	-17.7003	0.9	-52.1239	16.7234	FALSE
F3″	F5″	17.8049	0.9	-16.6188	52.2285	FALSE
F4″	F5″	35.5051	0.0357	1.0815	69.9288	TRUE

Therefore, the above statistical analyses on the vowel and word samples reveal that three formant transitions (F3″, F4″, and F5″) are not statistically significant features. The other two transitions (F1″ and F2″) are not significant in the vowel sample, but they appeared significant in the word sample. All other formant frequency-related features (five formants and five dispersions) are statistically significant.

3.3 Input Features and Hidden Layers Variation

Both vowel and word classification are performed by varying the number of input feature vectors and hidden layers. The classification model initializes its parameters randomly as discussed in Subsection 2.3.4, so four-fold cross-validation is utilized to accommodate the dependency of the model's performance on initialization parameters. The cross-validated (average) scores are shown in Table 3.8 with corresponding standard deviations in parentheses.

Table 3.8: Comparison between vowel and word classification performances by changing the number of feature vectors and hidden layers. The parameters represent the total number of trainable parameters. It is justified that the metrics for words have a lower value than that of vowels.

Features[1]	Layers[2]	Parameters	Type	Accuracy	AUC-ROC	F_1 Score	Cohen's κ
5F	HL_1	1,671	Vowel	0.76 ($\pm$ 0.00)	0.96 ($\pm$ 0.00)	0.76 ($\pm$ 0.00)	0.72 ($\pm$ 0.00)
			Word	0.50 ($\pm$ 0.01)	0.85 ($\pm$ 0.00)	0.50 ($\pm$ 0.01)	0.42 ($\pm$ 0.01)
	HL_2	9,479	Vowel	0.84 ($\pm$ 0.00)	0.98 ($\pm$ 0.00)	0.84 ($\pm$ 0.00)	0.81 ($\pm$ 0.00)
			Word	0.57 ($\pm$ 0.00)	0.89 ($\pm$ 0.00)	0.57 ($\pm$ 0.00)	0.50 ($\pm$ 0.00)
	HL_3	11,335	Vowel	0.86 ($\pm$ 0.01)	0.98 ($\pm$ 0.00)	0.86 ($\pm$ 0.01)	0.84 ($\pm$ 0.01)
			Word	0.61 ($\pm$ 0.01)	0.90 ($\pm$ 0.00)	0.60 ($\pm$ 0.00)	0.54 ($\pm$ 0.01)
	HL_4	11,751	Vowel	0.87 ($\pm$ 0.01)	0.98 ($\pm$ 0.00)	0.87 ($\pm$ 0.01)	0.85 ($\pm$ 0.01)
			Word	0.61 ($\pm$ 0.01)	0.90 ($\pm$ 0.00)	0.61 ($\pm$ 0.01)	0.54 ($\pm$ 0.01)
	HL_5	28,263	Vowel	0.89 ($\pm$ 0.00)	0.99 ($\pm$ 0.00)	0.89 ($\pm$ 0.00)	0.87 ($\pm$ 0.00)
			Word	0.64 ($\pm$ 0.01)	0.91 ($\pm$ 0.00)	0.63 ($\pm$ 0.01)	0.57 ($\pm$ 0.01)
	HL_6	32,423	Vowel	0.88 ($\pm$ 0.00)	0.98 ($\pm$ 0.00)	0.88 ($\pm$ 0.00)	0.86 ($\pm$ 0.00)
			Word	0.64 ($\pm$ 0.01)	0.91 ($\pm$ 0.00)	0.64 ($\pm$ 0.01)	0.58 ($\pm$ 0.01)
3F	HL_5	28,007	Vowel	0.88 ($\pm$ 0.01)	0.99 ($\pm$ 0.00)	0.88 ($\pm$ 0.01)	0.86 ($\pm$ 0.01)
			Word	0.64 ($\pm$ 0.01)	0.91 ($\pm$ 0.00)	0.64 ($\pm$ 0.01)	0.58 ($\pm$ 0.01)
5F, 5T	HL_5	28,903	Vowel	0.89 ($\pm$ 0.01)	0.98 ($\pm$ 0.00)	0.89 ($\pm$ 0.01)	0.87 ($\pm$ 0.01)
			Word	0.63 ($\pm$ 0.01)	0.90 ($\pm$ 0.00)	0.63 ($\pm$ 0.01)	0.57 ($\pm$ 0.01)
5F, 4D	HL_5	28,775	Vowel	0.89 ($\pm$ 0.00)	0.98 ($\pm$ 0.00)	0.89 ($\pm$ 0.00)	0.87 ($\pm$ 0.00)
			Word	0.64 ($\pm$ 0.01)	0.91 ($\pm$ 0.00)	0.64 ($\pm$ 0.01)	0.58 ($\pm$ 0.01)
5F, 5T, 4D	HL_5	29,415	Vowel	0.89 ($\pm$ 0.00)	0.98 ($\pm$ 0.00)	0.89 ($\pm$ 0.00)	0.87 ($\pm$ 0.01)
			Word	0.64 ($\pm$ 0.00)	0.90 ($\pm$ 0.00)	0.64 ($\pm$ 0.00)	0.58 ($\pm$ 0.00)

[1] 5F: Five formant frequency; 3F: First three formant frequency (F1, F2, F3); 5T: Five formant transitions; 4D: Four formant dispersions
[2] HL_1: One hidden layer having 128 neurons
[2] HL_2: Two hidden layers having 128 and 64 neurons, respectively
[2] HL_3: Three hidden layers having 128, 64, and 32 neurons, respectively
[2] HL_4: Four hidden layers having 128, 64, 32, and 16 neurons, respectively
[2] HL_5: Five hidden layers having 128, 128, 64, 32, and 16 neurons, respectively
[2] HL_6: Six hidden layers having 128, 128, 64, 64, 32, and 16 neurons, respectively

Table 3.8 reports both vowel and word classification scores starting from a model having only one hidden layer. Having five formants in the input layer, the vowel classification model gives an accuracy of 0.76, an AUC-ROC score of 0.96, and so on for other metrics. With

increasing the number of hidden layers one by one, we can see an increase of performance till the five-hidden-layered (HL_5) model. The low scores for models having less than five hidden layers indicate that the network was under-fitted (since increasing the number of hidden layers to five has increased the overall performance). The parameters of models with less than five hidden layers were insufficient to cover the acoustical variabilities completely. The highest scores observed for vowel classifications were from five hidden layers (HL_5) with five formant features in the input. Increasing the number of hidden layers to six did not improve the performance significantly for both vowel and word classification, which means the model parameters (of HL_5) are now adequate to overcome the acoustical variabilities.

Another thing to note is that the performance at five formant frequencies is very similar to that at the first three formant frequencies only. The word classification performance on three formant frequencies is even better. This phenomenon is explainable from Figure 3.4, where it is clearly visible that the variations of F4 and F5 are significantly less as compared to F1, F2, and F3. Smaller variations denote that they were more or less the same for all vowels or all words. Therefore, they did not contribute much to the classification of vowels or words. Removing F4 and F5 does not hugely reduce the number of parameters as this only reduces two input neurons. However, it appears than only three formant frequencies (F1, F2, F3) are sufficient for a satisfactory classification.

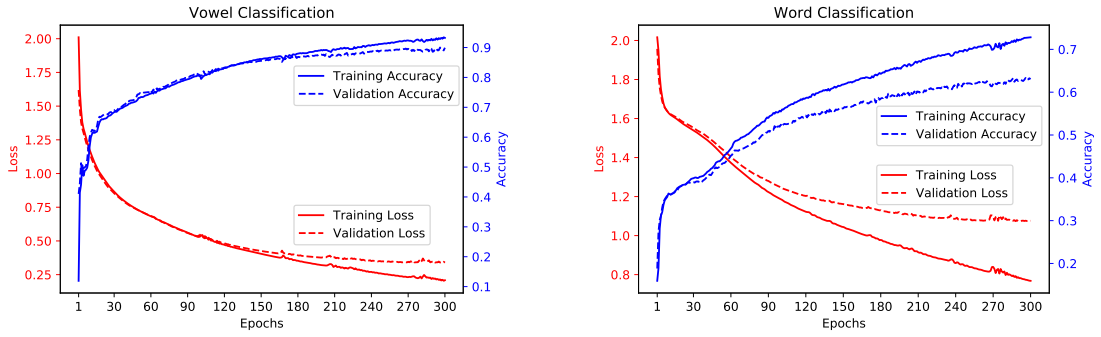
Furthermore, it can be concluded that the HL_5 configuration is an optimum choice here. In this case, the performance for vowel classification was 89% accuracy and 99% AUC-ROC, whereas, for word classification, these scores were 64% and 91%, respectively. It is worthwhile to mention that increasing the number of layers also increases the total number of trainable parameters. These additional parameters help the model to understand input representations in a simpler way. Nevertheless, here is a caveat that more hidden layers increase the computational burden, which is a crucial limit in implementing speech recognition systems in low-resource edge devices.

It is shown in the table that incorporating formant transitions and dispersions do not produce substantial benefits. For all four cases (all features, except dispersions, except transitions, and formants only), the performance scores are more or less equal to each other for the same five-hidden-layered (HL_5) configuration. Thus, formant transitions and dispersions are not that

essential features in the vowel or word classification. Therefore, on average, for any DNN configuration, word classification accuracy depends on the network's trainable parameters and poorly depends upon the transition and dispersion of formant. However, in canonical correlation-based studies, dispersions significantly influence classification [33], [34]. Our study proves that there is no impact of transitions and dispersions on speech classification. Therefore, the DNN optimization algorithm is compensating the necessity of including the formant transitions and dispersions. This result suggests that the hidden layers and neurons extract the underlying dispersion and transitional relationship from the five formant features, and thus, these dispersions and transitions are unnecessary to consider in classification. Hence, the research outcome is that the number of hidden layers should be extended instead of incorporating formant transitions and dispersions.

3.4 Graphical Comparison Between Vowel and Word Classification

From our previous discussions, we find five as an optimum number of hidden layers. Therefore, we have primarily utilized that HL_5 configuration to graphically illustrate training curves, validation curves, and performance comparisons between vowel and word classifications. The loss minimization and accuracy curves during both training and validation phases for both vowel and word classification are depicted in Figure 3.5. In this case, these figures came from only five formant frequencies as input vectors since previous experiments did not find any considerable importance of formant transitions and dispersions.

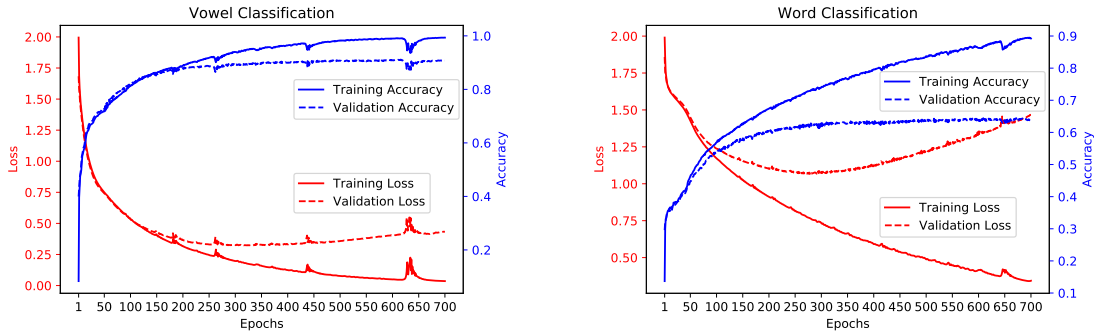


(a) Vowel classification. (b) Word classification.

Figure 3.5: Loss minimization and accuracy score during training and validation for 300 epochs. With increasing epochs, word's training and validation curves are being apart from each other, which denotes that word classification is more overfitting.

In general, DNNs tend to overfit training data, which can be verified by observing training and validation curves. If there is a considerable difference between training and validation curves, it is apparent that the model is suffering from the overfitting issue, which is also termed as noise-related high variance problem in deep learning literature. Here in our case for vowel classification, the training and validation curves have a good match, proving there are fewer higher variational acoustical features in vowels. On the contrary, words have higher variational acoustical features, as proved by the significant difference between training and validation curves. In short, lesser overfitting in vowel classification indicates fewer random acoustical features, whereas higher overfitting in word classification indicates higher random acoustical features. The vocal tract produces these random acoustical features during word utterances, and for this reason, the information of dynamic acoustic features is not adequately modeled like vowel classification.

Now, someone might ask why train up to 300 epochs only? In Figure 3.6, we can see that the validation curves are not improving (the losses are not decreasing, and the accuracies are not increasing) after 300 epochs. Instead, the validation performance deteriorates roughly after 300 epochs. Therefore, the model saturates at around 300 epochs. Training more than that resulted in higher overfitting in both vowel and word classification.



(a) Vowel classification. (b) Word classification.

Figure 3.6: Loss minimization and accuracy score during training and validation up to 700 epochs. After around 300 epochs, the validation scores did not improve. Rather, they worsen, which justifies training the model up to 300 epochs.

The validation phase's loss and accuracy curves for training up to 300 epochs have been compared between the vowel and the word classification (Figure 3.7).

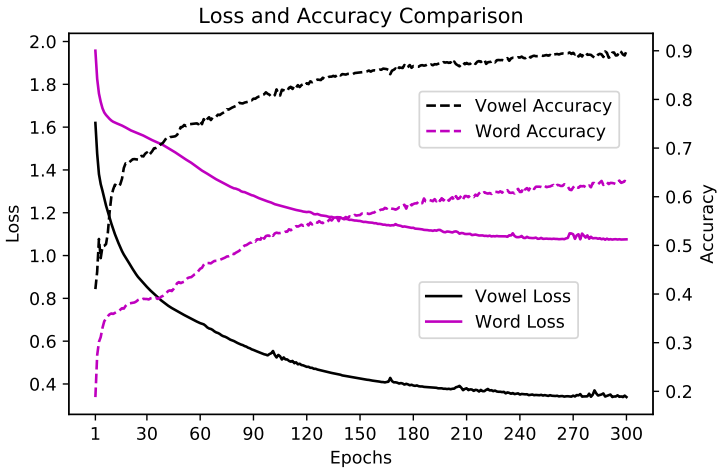


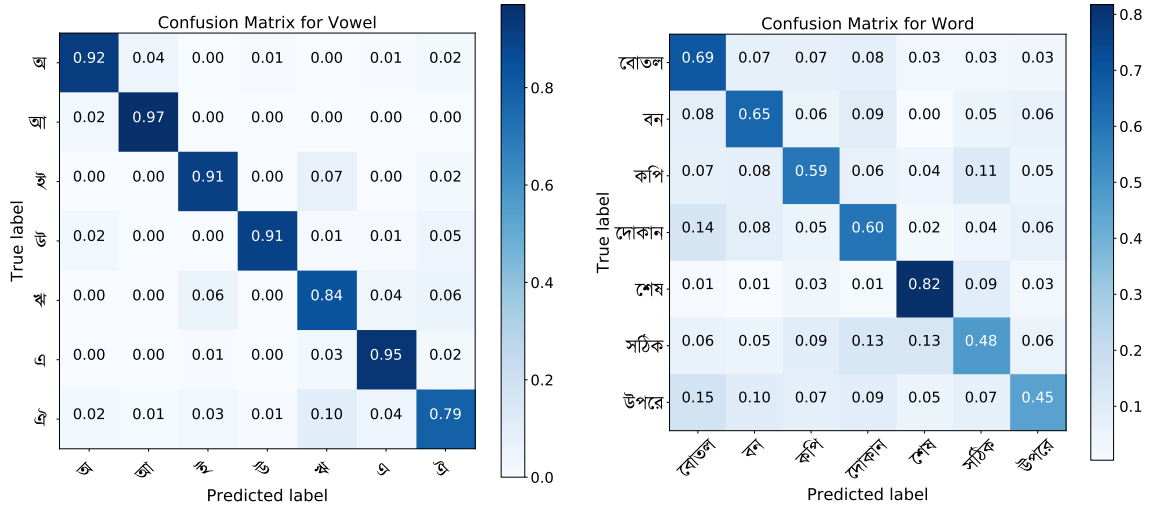
Figure 3.7: Validation loss and accuracy comparison between the vowel and the word classification. Vowel loss is optimized to a smaller value than the word loss, and the vowel accuracy is higher than the word accuracy.

After 300 epochs, the validation loss in vowel classification went down to 0.3388, whereas in word classification, it reached 1.0757. Similarly, the final validation accuracy in vowel classification was 89.94%, whereas, for words, it was only 63.05%. A lower loss indicates better cost function minimization, and thus, it further states that vowel classification is easier than word classification. From accuracy comparison, the higher accuracy denotes the higher rate of correct classification.

The confusion matrices are depicted in Figure 3.8, where top-left to bottom-right diagonal values represent the percentage of accurate classification for respective labels. If these values were all one, it would have been the best model.

As shown from the matrices that except for /ঋ/[/ri/] and /ঐ/[/oi/]) vowels, the model's correct prediction was above 90%. We observe the best performance (97% correct) for the vowel classification, particularly for /আ/[/a/] vowel. The vowel classification model's worst prediction accuracy (79%) was for /ঐ/[/oi/]) vowel. On the contrary, the word classification model's best performance (82% correct) was for the শেষ word. In all other words, the model's performance was comparatively lower than vowel classification. The worst performance (only 45% accurate) was found for the উপরে word.

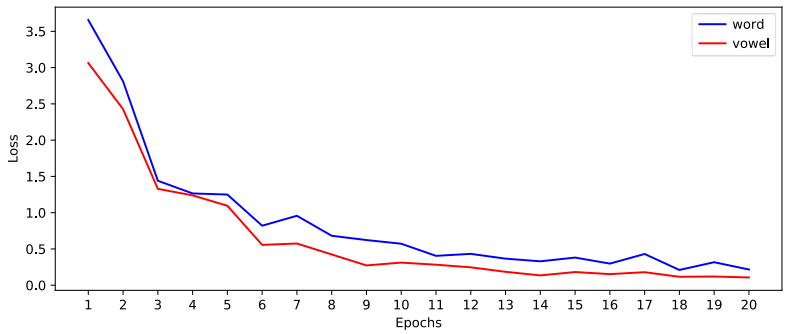
The classification performance of vowels and words is further reported using an advanced classifier (CNN) using MFCC features to examine how variance is absorbed.



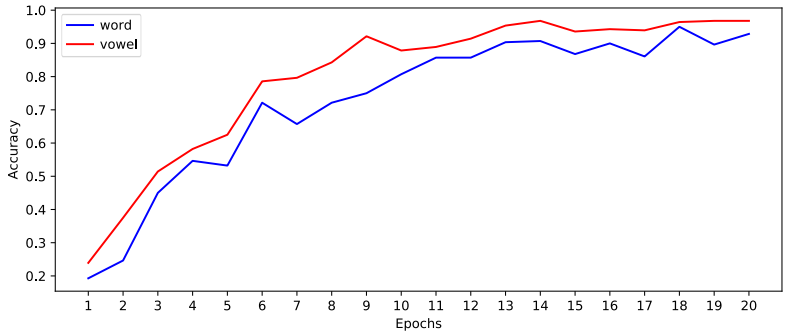
(a) Vowel classification.

(b) Word classification.

Figure 3.8: Confusion matrices for the vowel and the word classification using formant frequency features. It indicates that the vowel classification model is quite suitable for correctly classifying vowels, but the word classification model is more confused to classify words correctly.



(a) Loss comparison.



(b) Accuracy comparison.

Figure 3.9: Loss and accuracy comparison between vowel and word classification using a CNN classification model.

Figure 3.9a and Figure 3.9b depicts the loss and accuracy comparison between the vowel and word classification in a CNN classifier having one 2D convolutional layer (followed by a max-pooling, a dropout, and a flatten layer) and a dense hidden layer of 128 neurons (followed by another dropout layer) before the final output layer. The result presents a closer distance between vowel and word curves for both loss and accuracy metrics. Additionally, there is a significant increase in word classification accuracy when the CNN is utilized. Therefore as compared to the DNN model (comparison result in Figure 3.7), CNN provides better performance in terms of absorbing variance.

3.5 Summary

In all cases of Table 3.8, word classification significantly underperforms compared to vowel classification in all five different performance metrics. Additionally, word classification's deficiency is confirmed by comparing loss curves, accuracy curves (Figure 3.7), and confusion matrices (Figure 3.8). The reason behind such classification performance deviation lies in the variability of acoustical features. Vocal tract dynamics provides linguistic information related to dynamic acoustical features and random acoustic noises during word production. Additional parameterized DNNs are required to accommodate these dynamic acoustic features in classification and filter out individual articulation gestures of words. Thus, lower parameterized DNN poorly performs in word classification.

Some specific contributions explained in this chapter are as follows.

1. It demonstrates a process to relegate acoustical variability induced by vocal tract dynamics, which is a significant cause behind the performance degradation of spontaneous speech recognition.

2. It introduces a DNN-based framework to classify speech tokens and find the optimum number of hidden layers and input features. It further explain the classification pipeline with five different performance metrics.

3. It reports that formant transitions and dispersions have no important contribution to vowel and word classification, although canonical correlation-based studies found their significance.

Tripathi *et al.* [12] obtained 83% speech-mode classification accuracy with vocal tract features. It is evident that we should not strongly compare with this study because the datasets and classification domain are different. However, this comparison proves the applicability of our DNN-based classification model as we have achieved as high as 89% classification accuracy for vowel classification.

CHAPTER IV

Experimental Results: Optimum Number of MFCCs

This chapter reports and explains the findings on experiments to find out optimum number of MFCCs. In general, the optimum number of MFCC is first examined, then analysis between hidden layer increase and MFCC increase is presented, and finally, the best classification performance are extracted, followed by a comparison with related studies. A sketch of the steps involved in finding out optimum number of MFCCs is depicted in Figure 4.1.

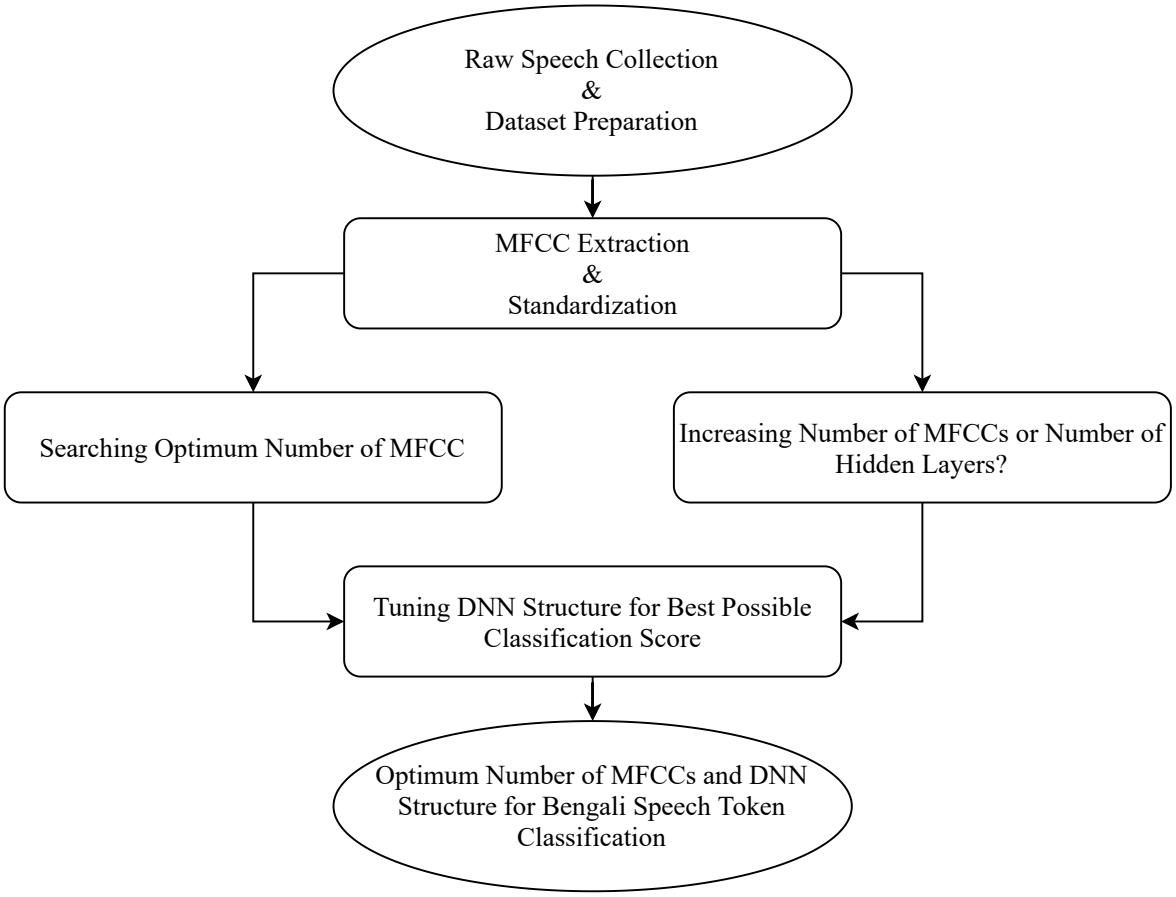


Figure 4.1: Steps involved in finding out optimum number of MFCCs.

This chapter incorporate three types of experiments (finding the optimum number of MFCC; suitability between MFCC and hidden layer increase; and securing the best possible score), and for these, the experimental setups are designed at first. For the classification of vowels and words, DNN architecture is utilized, and the structure is also tuned to obtain better classification results. Tuning involves the optimum number of hidden layers, number of neurons in them, activation functions of respective layers, loss function, optimizer and its learning rate, batch mode training, batch size, number of epochs, train-test split ratio, and evaluation metrics or accuracy indexes. In all these cases, there were multiple options, from where the optimum ones are selected by tracking the evaluation metrics. The descriptions of these steps are explained in Chapter II. The key steps involved in this chapter are mentioned below.

1. The datasets presented in Section 2.1 Description of the Datasets are processed according to Subsection 2.2.2 Mel-Frequency Cepstral Coefficient (MFCC), through which MFCCs are extracted. Then those features are standardized according to Subsection 2.3.1 Feature Standardization.

2. The standard features are then fed to the classification model described in Subsection 2.3.2 Model Configuration. The following three experiments are performed:

   (a) To seek the optimum number of MFCCs, the numbers of MFCC features are gradually increased from 8 to 28 for both vowel and word classification in the two-hidden-layered network presented in Figure 2.2.

   (b) As we have already discussed in Section 1.1 Background that both increasing the number of MFCCs and hidden layers raises the computational burden while enhancing the classification performance, this research also aims to find which increment is more suitable. For this purpose, the number of hidden layers are increased and corresponding classification metrics are calculated. Then both metrics (obtained by increasing the number of MFCCs and hidden layers) are compared based on the number of trainable parameters.

   (c) Finally, to elicit the best possible score, the optimum number of MFCCs found in Item 2a are utilized and then experimented by increasing the number of hidden layers and the number of neurons in them.

4.1 Search for Optimum Number of MFCC

The architecture depicted in Figure 2.2 is utilized for demonstrating performances for a varying number of MFCC for both vowel and word classification. Table 4.1 reports a detailed comparison with respect to the variation in the number of MFCC features for both vowel and the word classification.

Table 4.1: Performance comparison with respect to the variation of the number of MFCC features for both vowel and word classification. The parameters represent the total number of trainable parameters. It denotes that twenty-five MFCCs seems optimum for both vowel and word classification. For this whole comparison, the model is trained for 50 epochs in the DNN configuration of two hidden layers shown in Figure 2.2 with all random initialization fixed to a seed value of 42.

Type	MFCCs	Parameters	Accuracy	AUC-ROC	F_1 Score	Cohen's κ
Vowel	8	935	0.66 (± 0.01)	0.94 (± 0.00)	0.65 (± 0.01)	0.61 (± 0.01)
	10	999	0.70 (± 0.01)	0.95 (± 0.00)	0.68 (± 0.01)	0.64 (± 0.01)
	12	1063	0.71 (± 0.01)	0.95 (± 0.01)	0.71 (± 0.01)	0.67 (± 0.01)
	13	1095	0.74 (± 0.02)	0.96 (± 0.00)	0.73 (± 0.02)	0.69 (± 0.02)
	14	1127	0.75 (± 0.02)	0.96 (± 0.00)	0.74 (± 0.02)	0.71 (± 0.02)
	16	1191	0.78 (± 0.01)	0.97 (± 0.00)	0.78 (± 0.01)	0.74 (± 0.01)
	18	1255	0.80 (± 0.00)	0.97 (± 0.00)	0.80 (± 0.00)	0.77 (± 0.00)
	20	1319	0.82 (± 0.01)	0.98 (± 0.00)	0.82 (± 0.01)	0.79 (± 0.01)
	22	1383	0.83 (± 0.01)	0.98 (± 0.00)	0.82 (± 0.01)	0.80 (± 0.01)
	24	1447	0.84 (± 0.00)	0.98 (± 0.00)	0.84 (± 0.00)	0.81 (± 0.01)
	25	**1479**	**0.83 (± 0.01)**	**0.98 (± 0.00)**	**0.83 (± 0.01)**	**0.81 (± 0.01)**
	26	1511	0.83 (± 0.01)	0.98 (± 0.00)	0.82 (± 0.01)	0.80 (± 0.01)
	27	1543	0.84 (± 0.00)	0.98 (± 0.00)	0.84 (± 0.00)	0.81 (± 0.00)
	28	1575	0.84 (± 0.00)	0.98 (± 0.00)	0.84 (± 0.00)	0.82 (± 0.00)
Word	8	935	0.46 (± 0.00)	0.83 (± 0.00)	0.44 (± 0.00)	0.37 (± 0.00)
	10	999	0.47 (± 0.01)	0.84 (± 0.01)	0.45 (± 0.01)	0.38 (± 0.01)
	12	1063	0.49 (± 0.01)	0.85 (± 0.00)	0.47 (± 0.01)	0.40 (± 0.01)
	13	1095	0.51 (± 0.01)	0.86 (± 0.00)	0.49 (± 0.01)	0.42 (± 0.01)
	14	1127	0.52 (± 0.01)	0.86 (± 0.00)	0.51 (± 0.01)	0.44 (± 0.01)
	16	1191	0.51 (± 0.00)	0.86 (± 0.00)	0.50 (± 0.00)	0.43 (± 0.01)
	18	1255	0.53 (± 0.01)	0.87 (± 0.00)	0.52 (± 0.01)	0.45 (± 0.01)
	20	1319	0.54 (± 0.01)	0.87 (± 0.00)	0.52 (± 0.01)	0.46 (± 0.01)
	22	1383	0.55 (± 0.00)	0.88 (± 0.00)	0.54 (± 0.00)	0.47 (± 0.00)
	24	1447	0.56 (± 0.01)	0.88 (± 0.00)	0.55 (± 0.01)	0.48 (± 0.01)
	25	**1479**	**0.57 (± 0.01)**	**0.88 (± 0.00)**	**0.56 (± 0.01)**	**0.49 (± 0.01)**
	26	1511	0.56 (± 0.01)	0.88 (± 0.00)	0.55 (± 0.01)	0.48 (± 0.01)
	27	1543	0.57 (± 0.01)	0.88 (± 0.00)	0.56 (± 0.01)	0.49 (± 0.01)
	28	1575	0.56 (± 0.00)	0.88 (± 0.00)	0.56 (± 0.00)	0.49 (± 0.00)

Since the performance of the model may vary at different runtimes, four-fold cross-validation is utilized, and the average scores of these four runtimes are shown in Table 4.1 along with the standard deviation in parentheses. While comparing, all random initializations are fixed

from a constant seed value of 42. This technique ensures the reproducibility of identical performance scores in multiple runtimes. For this comparison table, the two-hidden-layered network was trained for 50 epochs.

For both vowel and word classification, the performance scores increase with an increasing number of MFCCs. For vowel classification, they were highest for 24 MFCCs in input, whereas for word classification, the highest scores of the same metrics were observed for 25 MFCCs in input. From 13 MFCCs to 25 MFCCs, vowel shows $0.83 - 0.74 = 0.09$ i.e. 9%, and word shows $0.57 - 0.51 = 0.06$ i.e. 6% increase in overall accuracy. It is seen from Table 4.1 that 24 MFCCs takes less parameters and provides better accuracy and F_1 score for vowels, and provides lower accuracy, F_1 score, and Cohen's κ for words compared to 25 MFCCs, although the difference is much smaller. One can use either 24 or 25 MFCCs for their evaluation, but 25 MFCCs are chosen for being consistent.

4.2 MFCC Increase vs. Hidden Layer Increase

It is apparent that increasing the number of hidden layers should also increase the performance, and that is why the same classification experiment is performed with three hidden layers (Table 4.2), and as expected, the performance increased. However, the use of three hidden layers increases the number of trainable parameters, thereby increasing computational burden, which is a crucial limit in implementing speech recognition systems in low-resource devices, such as microcontrollers and Arduino.

Table 4.2: Performance comparison with increasing the number of hidden layers at 13 MFCCs. The model was cross-validated (4-fold) and trained for 50 epochs.

Type	Model	Parameters	Time (sec)[c]	Accuracy	AUC-ROC	F_1 Score	Cohen's κ
Vowel	HL_2[a]	1095	5.40	0.74 ($\pm$ 0.02)	0.96 ($\pm$ 0.00)	0.73 ($\pm$ 0.02)	0.69 ($\pm$ 0.02)
	HL_3[b]	3623	5.88	0.79 ($\pm$ 0.01)	0.97 ($\pm$ 0.00)	0.79 ($\pm$ 0.01)	0.76 ($\pm$ 0.01)
Word	HL_2	1095	5.83	0.51 ($\pm$ 0.01)	0.86 ($\pm$ 0.00)	0.49 ($\pm$ 0.01)	0.42 ($\pm$ 0.01)
	HL_3	3623	6.29	0.55 ($\pm$ 0.01)	0.89 ($\pm$ 0.00)	0.55 ($\pm$ 0.01)	0.48 ($\pm$ 0.01)

[a] HL_2: Two hidden layers having 32 and 16 neurons, respectively
[b] HL_3: Three hidden layers having 64, 32, and 16 neurons, respectively
[c] Time (sec): The execution time (in seconds) of both the training and validation phase together.

Table 4.2 also reports total execution time (both training and validation phase together) in an HP Pavilion 14 laptop with 64-bit Linux Mint 19.3 OS, 8 GB RAM, 1.7 GHz Intel Core

i5-4210U Processor, and NVIDIA GM108M [GeForce 830M] 2 GB Graphics. As shown in the table, for vowel classification employing 13 MFCCs, two hidden layers to three hidden layers introduce 5% increase in accuracy, 1% increase in AUC-ROC, 6% increase in F_1 Score, and 7% increase in Cohen's κ. At the same time, these performance increases require $3623 - 1095 = 2528$ additional trainable parameters. Similar increases of performances in three hidden layers is comparable to 16 or 17 MFCCs of Table 4.1. Therefore, if we increase the number of MFCCs to increase the performances, it would cost only $1191 - 1095 = 96$ additional trainable parameters. As compared to execution time, 16 MFCCs required only 5.53 seconds which is also less than the execution time in three-hidden-layered configuration (HL_3). Comparing these additional parameters in two cases, we can comment that increasing MFCCs is computationally efficient than increasing hidden layers since more trainable parameters require more time and computational power to do the classification. Word classification also depicts similar results. An increase of 4% classification accuracy can be done either by additional 2528 parameters if we want to increase the number of hidden layers or by additional $1383 - 1095 = 288$ parameters if we want to increase the number of MFCCs to 22. As for execution time, 22 MFCCs took only 5.75 seconds, less than what it took for HL_3. Accordingly, this guides us to increase the number of MFCCs rather than the number of hidden layers. The analysis further suggests that the optimum number of MFCCs could be 24 or 25, although the optimum number of MFCCs is well-known to be 13.

In a search for the optimum number of MFCCs in English and Portuguese digit classification, Silva *et al.* [9] reported a decrease in performance after 25 MFCCs, and they concluded that the optimum number of coefficients has a narrow range between 11 to 23 MFCCs. Our findings also comply with it. Authors in [10] also experimented by varying the number of MFCCs and other parameters to find the best configuration for a low-resource embedded system. They experimented with 8, 9, and 10 MFCCs, and concluded that 9 MFCCS are suitable in accordance with other parameter tuning such as the number of filters and number of HMM states. A significant reason behind choosing 9 MFCCs was less training and recognition time, particularly for low-resource embedded systems they considered.

4.3 Search for Best Scores

The classification performances have a proportional relationship with the number of MFCCs and the number of hidden layers. According to the previous discussions (Section 4.2 MFCC Increase vs. Hidden Layer Increase), to enhance the classification score, we have to first concentrate on incorporating the optimum number of MFCC, and then, we should extend the number of hidden layers. As we already concluded that 25 MFCCs would be the optimum number of MFCCs, the number of hidden layers is increased while utilizing 25 MFCCs shown in Table 4.3.

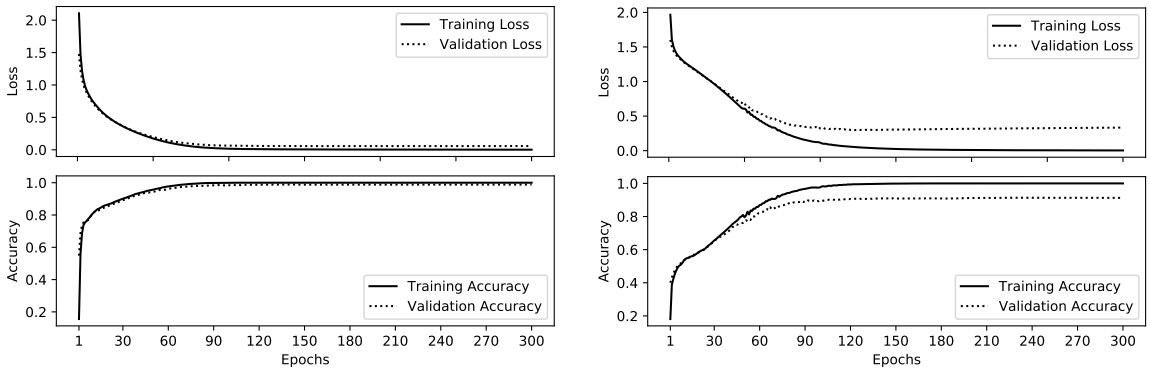
Table 4.3: Best performance scores with increasing the number of hidden layers at 25 MFCCs. The evaluation metrics are 4-fold cross-validated.

Type	Model	Accuracy	AUC-ROC	F_1 Score	Cohen's κ
Vowel	HL_2[a]	0.96 ($\pm$ 0.00)	1.00 ($\pm$ 0.00)	0.96 ($\pm$ 0.00)	0.95 ($\pm$ 0.00)
	HL_4[b]	0.99 ($\pm$ 0.00)	1.00 ($\pm$ 0.00)	0.99 ($\pm$ 0.00)	0.98 ($\pm$ 0.00)
	HL_5[c]	**0.99 ($\pm$ 0.00)**	**1.00 ($\pm$ 0.00)**	**0.99 ($\pm$ 0.00)**	**0.98 ($\pm$ 0.00)**
Word	HL_2	0.75 ($\pm$ 0.01)	0.96 ($\pm$ 0.00)	0.75 ($\pm$ 0.01)	0.71 ($\pm$ 0.01)
	HL_4	0.89 ($\pm$ 0.01)	0.98 ($\pm$ 0.00)	0.89 ($\pm$ 0.01)	0.87 ($\pm$ 0.01)
	HL_5	**0.91 ($\pm$ 0.01)**	**0.98 ($\pm$ 0.00)**	**0.91 ($\pm$ 0.01)**	**0.90 ($\pm$ 0.01)**

[a] HL_2: Two hidden layers having 32 and 16 neurons, respectively
[b] HL_4: Four hidden layers having 128, 64, 32, and 16 neurons, respectively
[c] HL_5: Five hidden layers having 128, 128, 64, 32, and 16 neurons, respectively

While increasing the number of hidden layers, five hidden layers are found as optimum since performance scores did not increase for six and more hidden layer configurations. A varying number of hidden neurons is checked in the configurations shown in Table 4.3, but the reported configurations were found optimum. Finally, the model is trained for 300 epochs to extract the best possible scores, which resulted in 99% and 91% classification accuracies for vowels and words, respectively.

For the best scores, Figure 4.2a and Figure 4.2b depict the loss minimization and accuracy score curves during both training and validation phases for both vowel and word classification, respectively. Since there is no large difference between training and validation curves for vowel classification, the model is not overfitting.

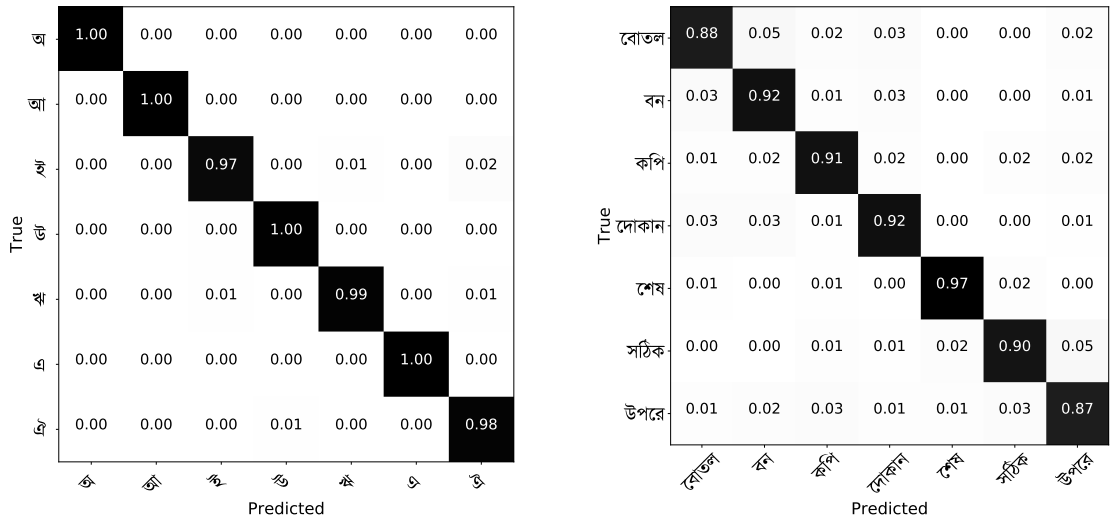


(a) Vowel classification. (b) Word classification.

Figure 4.2: Loss minimization and accuracy score during training and validating using MFCC features. The classification model consists of five hidden layers.

The smoother curves for vowel classification (Figure 4.2a) indicate proper tuning of model hyperparameters. It further means that the formation of the DNN classifier is good enough to have such a good model. However, for word classification, there are some differences between training and validation losses shown in Figure 4.2b. A possible reason behind this is the presence of more dynamic acoustical features in words [89].

Classification performance for individual labels can be identified from the confusion matrices shown in Figure 4.3. The top-left to bottom-right diagonal elements in the matrix represent the accurate classification rate for respective classes.



(a) Vowel classification. (b) Word classification.

Figure 4.3: Confusion matrices for the vowel and the word classification using MFCC features. It indicate the classification performances of individual vowels and words.

As shown from the matrices, vowel classification was perfect with 100% accuracy for /অ/[/ɔ/], /আ/[/a/], /উ/[/u/], and /এ/[/e/] vowels. The class that had the lowest accuracy with this model was /ই/[/i/], and it was 97% correct. On the contrary, the model's best performance for the word classification was for the শেষ word, and it was 97% correct. For all other words, the model's performance was quite satisfactory with an 87% classification rate as the lowest found for উপরে word.

4.4 Literature Comparison Based on Number of MFCCs Used

Our main target here is to investigate the number of MFCC features required to achieve the best classification performance. Table 4.4 reports some relevant studies, where at first, the DNN-based studies are presented, then the CNN-based studies followed by other approaches are presented. In each category, the articles are ordered according to the highest classification performance to the lowest.

Table 4.4 reveals that most of the speech classification researches employ either DNN or CNN, which is the extended version of DNN architecture added with convolutional layers. Several studies also utilize the 1^{st} and 2^{nd} derivatives of 13 base MFCCs to consider the temporal dynamics of the speech signal [39]. Since the recognition domain and datasets are different, we should not strictly compare the classification performance among these studies. However, competitive classification performance proves the applicability and suitability of our model. There are two concrete outcomes from this table—most articles utilized only 13 base MFCCs where our research recommends 24 or 25 MFCCs, and our classification accuracy outperforms state-of-the-art scores on vowel recognition. Also, our word classification accuracy is competitive to similar studies.

Table 4.4: Comparison with relevant studies based on the used number of MFCCs and reported classification accuracy. The comparison proves the competitiveness of our classification model with 25 MFCCs, whereas the general wisdom is to use 13 MFCCs.

Model	Article	Recognition Domain	MFCCs	Accuracy
DNN	**Proposed**	**Seven Bengali vowels**	**25**	**99%**
	Wahyuni [46]	Three Arabic words	13	92.42%
	Proposed	**Seven Bengali words**	**25**	**91%**
	Bird *et al.* [43]	English phonemes	Not Specified	90.77%
	Yang *et al.* [13]	English command words	13	82.46%
	Syfullah *et al.* [20]	Bengali speech characters	Not Specified	81.61%
	Mohamed and Lajish [45]	Malayalam Vowels	12	74.39%
CNN	Sharmin *et al.* [15]	10 spoken Bengali digits	Not Specified	98.37%
	Soliman *et al.* [47]	Isolated English words	Not Specified	96.19%
	Salau *et al.* [3]	Nigerian accent classification	40	94.9%
	Hasan and Hasan [89]	Isolated Bengali vowels	13	93.93%
	Hasan and Hasan [89]	Isolated Bengali words	13	90 %
	Dawodi *et al.* [14]	Dari speech tokens	13	88.2%
	Islam *et al.* [18]	Isolated Bengali speech	Not Specified	86.058%
	Sumon *et al.* [21]	Ten Bengali short speech words	13	74.01%
SVM	Cen *et al.* [48]	Speech emotion	13, Δ, $\Delta - \Delta$	90%
CMUSphinx	Das *et al.* [16]	Bengali word	13, Δ, $\Delta - \Delta$	85.3%
	Mandal *et al.* [17]	Bengali speech	13, Δ, $\Delta - \Delta$	Not Specified
HTK	Das *et al.* [16]	Bengali phoneme	13, Δ, $\Delta - \Delta$	54.07%

DNN: Deep Neural Network
CNN: Convolutional Neural Network
SVM: Support Vector Machine
CMUSphinx: Open source speech recognition system developed at Carnegie Mellon University
HTK: Hidden Markov Model Toolkit
Δ: 1st derivatives of 13 base MFCCs
$\Delta - \Delta$: 2nd derivatives of 13 base MFCCs

CHAPTER V

Conclusions and Future Works

Speech is a significant natural source of information used in many aspects—speech dictation gadgets, accent classification, emotion recognition, and disease diagnosis, to name a few. Therefore, speech-related researches have a meaningful impact on our day-to-day life. Since audio data cannot be processed directly in many cases, researchers extract valuable information from the audio, what we call feature extraction. Formant frequency and MFCC were chosen as acoustical features for the classification task since they have been widely used by researchers worldwide. The primary focus of this study is to investigate the effect of vocal tract dynamics on DNN-based speech recognition using formant frequency. For this, two separate datasets for seven Bengali vowels and seven Bengali words are utilized. Furthermore, two additional features—formant transitions and dispersions, were also derived from formant trajectory and used in the classification. As for the classification model, a feedforward neural network (DNN) was optimized by proper hyperparameter tuning. Five well-known performance metrics—*classification accuracy, AUC-ROC, F_1 score, Cohen's Kappa (κ)*, and *confusion matrix* were utilized to determine the classification performance. Since the classifications were also performed by varying the number of hidden layers and input features, this thesis also identified how many hidden layers and which set of features contribute more to the classification. This research found that formant transitions and dispersions have not introduced any added benefits in classification, and the DNN configuration having five hidden layers was an optimum choice. Furthermore, statistical analyses using ANOVA and Tukey's HSD tests reveal that formant transitions are not statistically significant as well. Specifically, the third, fourth, and fifth formant transitions are not statistically significant at all. Word classification performance lagged behind vowel classification by a large margin in all five metrics in all different tests. Initially, the vocal tract dynamics is checked from speech waveshapes, formant trajectories, and coefficient of variations of formant frequency.

From these, it was verified that words consist of more acoustic variability than vowels. Also, it is well known that during vowel pronunciation, our vocal tract becomes relatively steady, whereas, for word pronunciation, the vocal tract changes quite rapidly due to coarticulation, which eventually induces acoustical feature variations. Thus, the variation in words produced by vocal tract dynamics has lowered the classification performance. However, the amount of classification performance deviations are not quantified in this study. Future research might focus on relating these vocal tract dynamics and classification performance deviations quantitatively.

The secondary target of this thesis is to find out the optimum number of MFCC features for a satisfactory classification. Although it is a general wisdom to use the first 13 coefficients, we put a test to answer the question—how many MFCCs are to be utilized? The same classification of seven Bengali vowels and seven Bengali words is performed by varying MFCC numbers from 8 to 28. In a two-hidden-layered DNN model, 13 MFCCs gave 74% vowel and 51% word classification accuracy, whereas 25 MFCCs gave 83% vowel and 57% word classification accuracy, from which this thesis recommends that 25 MFCCs could be the optimum number of MFCCs. In fact, the general wisdom of 13 MFCCs could serve the same performance score if we increase the number of DNN hidden layers. However, it increases the total trainable parameters (computational burden), a limiting factor to implement speech recognition systems in edge devices like Arduino. With the optimum number of MFCCs discovered in this study, this research further seeks the best possible scores by increasing the hidden layers. Accordingly, in a five-hidden-layered model, this thesis obtains 99% vowel and 91% word classification accuracy that is competitive to other similar speech classification studies.

There are several unique applications of these findings. Such a formant frequency and DNN-based speech classifier can be employed as an acoustic to sound mapping tool in computational speech motor movement models. Neurological diseases such as Parkinson's disease, Dementia, Alzheimer's disease, and several other diseases like dysphagia, cleft palate, and oral cancer induce acoustical variability in the vocal tract. Neurological disorders such as Parkinson's and Alzheimer's disease can be diagnosed by comparing patients' acoustic variabilities with these (regular Bengali speakers) obtained in this article. Accordingly, such disease diagnosis schemes might benefit people only speaking the Bengali language. Also,

the vocal tract dynamics needs to be considered in spontaneous conversational speech recognition since one of the serious reasons behind the lack of progress in spontaneous speech recognition is this variability. Furthermore, the outcome of this study will be helpful for future MFCC-based researches. Employing the optimum number of MFCCs should increase the overall performance of devices, systems, and research involving MFCC. Apart from these, proper detection of these isolated speech tokens has several other applications, including speech dictation gadgets and services, emotion recognition, accent detection, and assisting physically-challenged and old-age people. Accordingly, service providers of these gadgets and services shall serve the Bengali language to their end-users with the aid of this Bengali speech token classification and the datasets. The information found in this study, particularly formant transitions and dispersions' no significant contribution to classification, may be utilized in feature selections in future researches in this domain. Thus, the outcomes of this research will help better design speech recognition-based devices and systems.

REFERENCES

[1] M. Yağanoğlu, "Real time wearable speech recognition system for deaf persons," *Computers & Electrical Engineering*, vol. 91, p. 107 026, 2021, ISSN: 0045-7906. DOI: 10.1016/j.compeleceng.2021.107026.

[2] A. Veiga, C. Lopes, L. Sá, and F. Perdigão, "Acoustic similarity scores for keyword spotting," in *Computational Processing of the Portuguese Language*, J. Baptista, N. Mamede, S. Candeias, I. Paraboni, T. A. S. Pardo, and M. d. G. Volpe Nunes, Eds., Springer, Cham, 2014, pp. 48–58. DOI: 10.1007/978-3-319-09761-9_5.

[3] A. O. Salau, T. D. Olowoyo, and S. O. Akinola, "Accent classification of the three major nigerian indigenous languages using 1d cnn lstm network model," in *Advances in Computational Intelligence Techniques*, S. Jain, M. Sood, and S. Paul, Eds., Singapore: Springer, 2020, pp. 1–16, ISBN: 978-981-15-2620-6. DOI: 10.1007/978-981-15-2620-6_1.

[4] M. J. S. Suresh and S. Thorat, "Language identification system using mfcc and sdc feature," in *Proceedings of 4th RIT Post Graduates Conference (RIT PG Con-18)*, Novateur Publication, 2018, pp. 113–119.

[5] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics," *The Journal of the Acoustical Society of America*, vol. 108, no. 6, pp. 3036–3048, 2000. DOI: 10.1121/1.1315288.

[6] Y.-C. Deng, C.-H. Lin, Y.-F. Liao, Y.-R. Wang, and S.-H. Chen, "Prosodic information-assisted dnn-based mandarin spontaneous-speech recognition," in *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2020, pp. 134–138. DOI: 10.1109/O-COCOSDA50338.2020.9295010.

[7] S. E. G. Öhman, "Coarticulation in vcv utterances: Spectrographic measurements," *The Journal of the Acoustical Society of America*, vol. 39, no. 1, pp. 151–168, 1966. DOI: 10.1121/1.1909864.

[8] S. E. G. Öhman, "Numerical model of coarticulation," *The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 310–320, 1967. DOI: 10.1121/1.1910340.

[9] D. F. Silva, V. M. A. d. Souza, and G. E. A. P. A. Batista, "A comparative study between mfcc and lsf coefficients in automatic recognition of isolated digits pronounced in portuguese and english," *Acta Scientiarum. Technology*, vol. 35, no. 4, pp. 621–628, May 2013. DOI: 10.4025/actascitechnol.v35i4.19825.

[10] D. D. C. da Silva, C. R. Vasconcelos, B. G. A. Neto, and J. M. Fechine, "Evaluation of the impact in reducing the number of parameters for continuous speech recognition for brazilian portuguese," in *2012 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC)*, 2012, pp. 1–6. DOI: 10.1109/BRC.2012.6222182.

[11] D. M. Eberhard, G. F. Simons, and C. D. Fennig, Eds., *Ethnologue: Languages of the World*, 24th ed. Dallas, Texas: SIL International, 2021. [Online]. Available: https://www.ethnologue.com.

[12] K. Tripathi, M. K. Reddy, and K. S. Rao, "Multilingual and multimode phone recognition system for indian languages," *Speech Communication*, vol. 119, pp. 12–23, 2020, ISSN: 0167-6393. DOI: 10.1016/j.specom.2020.02.006.

[13] X. Yang, H. Yu, and L. Jia, "Speech recognition of command words based on convolutional neural network," in *2020 International Conference on Computer Information and Big Data Applications (CIBDA)*, 2020, pp. 465–469. DOI: 10.1109/CIBDA50819.2020.00110.

[14] M. Dawodi, J. A. Baktash, T. Wada, N. Alam, and M. Z. Joya, "Dari speech classification using deep convolutional neural network," in *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2020, pp. 1–4. DOI: 10.1109/IEMTRONICS51293.2020.9216370.

[15] R. Sharmin, S. K. Rahut, and M. R. Huq, "Bengali spoken digit classification: A deep learning approach using convolutional neural network," *Procedia Computer Science*, vol. 171, pp. 1381–1388, 2020, ISSN: 1877-0509. DOI: 10.1016/j.procs.2020.04.148.

[16] B. Das, S. Mandal, and P. Mitra, "Bengali speech corpus for continuous auutomatic speech recognition system," in *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, 2011, pp. 51–55. DOI: 10.1109/ICSDA.2011.6085979.

[17] S. Mandal, B. Das, and P. Mitra, "Shruti-ii: A vernacular speech recognition system in bengali and an application for visually impaired community," in *2010 IEEE Students Technology Symposium (TechSym)*, 2010, pp. 229–233. DOI: 10.1109/TECHSYM.2010.5469156.

[18] J. Islam, M. Mubassira, M. R. Islam, and A. K. Das, "A speech recognition system for bengali language using recurrent neural network," in *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, 2019, pp. 73–76. DOI: 10.1109/CCOMS.2019.8821629.

[19] S. M. S. I. Badhon, M. H. Rahaman, F. R. Rupon, and S. Abujar, "State of art research in bengali speech recognition," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, pp. 1–6. DOI: 10.1109/ICCCNT49239.2020.9225650.

[20] S. M. Syfullah, Z. B. Zakaria, M. P. Uddin, M. F. Rabbi, M. I. Afjal, and A. M. Nitu, "Efficient vector code-book generation using k-means and linde-buzo-gray (lbg) algorithm for bengali voice recognition," in *2018 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, 2018, pp. 1–4. DOI: 10.1109/ICAEEE.2018.8642994.

[21] S. A. Sumon, J. Chowdhury, S. Debnath, N. Mohammed, and S. Momen, "Bangla short speech commands recognition using convolutional neural networks," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018, pp. 1–6. DOI: 10.1109/ICBSLP.2018.8554395.

[22] H. Mukherjee, S. Phadikar, and K. Roy, "An ensemble learning-based bangla phoneme recognition system using lpcc-2 features," in *Intelligent Engineering Informatics*, V. Bhateja, C. A. Coello Coello, S. C. Satapathy, and P. K. Pattnaik, Eds., Singapore: Springer Singapore, 2018, pp. 61–69. DOI: 10.1007/978-981-10-7566-7_7.

[23] S. Young, "Hmms and related speech recognition technologies," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 539–558, ISBN: 978-3-540-49127-9. DOI: 10.1007/978-3-540-49127-9_27.

[24] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012. DOI: 10.1109/MSP.2012.2205597.

[25] E. Trentin and M. Gori, "Robust combination of neural networks and hidden markov models for speech recognition," *IEEE Transactions on Neural Networks*, vol. 14, no. 6, pp. 1519–1531, 2003. DOI: 10.1109/TNN.2003.820838.

[26] C. M. Bishop, *Pattern Recognition and Machine Learning*. USA: Springer-Verlag New York, 2006, ISBN: 978-0387-31073-2.

[27] T. S. Shanthi and C. Lingam, "Review of feature extraction techniques in automatic speech recognition," *International Journal of Scientific Engineering and Technology*, vol. 2, no. 6, pp. 479–484, 2013.

[28] S. Gaikwad, B. Gawali, P. Yannawar, and S. Mehrotra, "Feature extraction using fusion mfcc for continuous marathi speech recognition," in *2011 Annual IEEE India Conference*, 2011, pp. 1–5. DOI: 10.1109/INDCON.2011.6139372.

[29] G. Fant, *Acoustic theory of speech production*, 2. Walter de Gruyter, 1970.

[30] R. D. Kent and C. Read, *The acoustic analysis of speech*, 2nd. USA: Singular Publishing Group San Diego, 2001, ISBN: 978-0-76-930112-9.

[31] B. H. Story and K. Bunton, "Relation of vocal tract shape, formant transitions, and stop consonant identification," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 6, pp. 1514–1528, Dec. 2010. DOI: 10.1044/1092-4388(2010/09-0127).

[32] J. D. Stephens and L. L. Holt, "A standard set of american-english voiced stop-consonant stimuli from morphed natural speech," *Speech Communication*, vol. 53, no. 6, pp. 877–888, 2011, ISSN: 0167-6393. DOI: 10.1016/j.specom.2011.02.007.

[33] S. López, P. Riera, M. F. Assaneo, M. Eguía, M. Sigman, and M. A. Trevisan, "Vocal caricatures reveal signatures of speaker identity," *Scientific reports*, vol. 3, no. 3407, pp. 1–7, 2013. DOI: 10.1038/srep03407.

[34] M. M. Hasan, S. R. Mitra, and K. Teramoto, "Canonical correlation based impersonation quality determination algorithm for natural morphed speech," in *2015 IEEE International Conference on Telecommunications and Photonics (ICTP)*, 2015, pp. 1–4. DOI: 10.1109/ICTP.2015.7427943.

[35] S. A. M. Yusof, P. M, and S. Yaacob, "Classification of malaysian vowels using formant based features," *Journal of Information and Communication Technology*, vol. 7, pp. 27–40, 2008, ISSN: 2180-3862. [Online]. Available: http://e-journal.uum.edu.my/index.php/jict/article/view/8076.

[36] J. Hillenbrand and R. T. Gayvert, "Vowel classification based on fundamental frequency and formant frequencies," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 4, pp. 694–700, 1993. DOI: 10.1044/jshr.3604.694.

[37] V. Vuckovic and M. Stankovic, "Formant analysis and vowel classification methods," in *5th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Service. TELSIKS 2001. Proceedings of Papers (Cat. No.01EX517)*, vol. 1, 2001, pp. 21–24. DOI: 10.1109/TELSKS.2001.954841.

[38] Q. Yan and S. Vaseghi, "Analysis, modelling and synthesis of formants of british, american and australian accents," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, vol. 1, 2003, pp. 712–715. DOI: 10.1109/ICASSP.2003.1198880.

[39] K. S. Rao and K. E. Manjunath, *Speech recognition using articulatory and excitation source features*. Springer, Cham, 2017, ISBN: 978-3-31-949220-9. DOI: 10.1007/978-3-319-49220-9.

[40] S. Sultana, M. S. Rahman, and M. Z. Iqbal, "Recent advancement in speech recognition for bangla: A survey," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 3, pp. 546–552, 2021. DOI: 10.14569/IJACSA.2021.0120365.

[41] F. Al-Anzi and D. AbuZeina, "Literature survey of arabic speech recognition," in *2018 International Conference on Computing Sciences and Engineering (ICCSE)*, 2018, pp. 1–6. DOI: 10.1109/ICCSE1.2018.8374215.

[42] K. J. Devi, N. H. Singh, and K. Thongam, "Automatic speaker recognition from speech signals using self organizing feature map and hybrid neural network," *Microprocessors and Microsystems*, vol. 79, p. 103 264, 2020, ISSN: 0141-9331. DOI: 10.1016/j.micpro.2020.103264.

[43] J. J. Bird, E. Wanner, A. Ekárt, and D. R. Faria, "Optimisation of phonetic aware speech recognition through multi-objective evolutionary algorithms," *Expert Systems with Applications*, vol. 153, p. 113 402, 2020, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2020.113402.

[44] S. Shahriar and Y. Kim, "Audio-visual emotion forecasting: Characterizing and predicting future emotion using deep learning," in *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, 2019, pp. 1–7. DOI: 10.1109/FG.2019.8756599.

[45] F. K. Mohamed and V. Lajish, "Nonlinear speech analysis and modeling for malayalam vowel recognition," *Procedia Computer Science*, vol. 93, pp. 676–682, 2016, Proceedings of the 6th International Conference on Advances in Computing and Communications, ISSN: 1877-0509. DOI: 10.1016/j.procs.2016.07.261.

[46] E. S. Wahyuni, "Arabic speech recognition using mfcc feature extraction and ann classification," in *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2017, pp. 22–25. DOI: 10.1109/ICITISEE.2017.8285499.

[47] A. Soliman, S. Mohamed, and I. A. Abdelrahman, "Isolated word speech recognition using convolutional neural network," in *2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, 2021, pp. 1–6. DOI: 10.1109/ICCCEEE49695.2021.9429684.

[48] L. Cen, F. Wu, Z. L. Yu, and F. Hu, "Chapter 2 - a real-time speech emotion recognition system and its application in online learning," in *Emotions, Technology, Design, and Learning*, ser. Emotions and Technology, S. Y. Tettegah and M. Gartmeier, Eds., San Diego: Academic Press, 2016, pp. 27–46, ISBN: 978-0-12-801856-9. DOI: 10.1016/B978-0-12-801856-9.00002-5.

[49] B. Kühnert and F. Nolan, "The origin of coarticulation," in *Coarticulation: Theory, Data and Techniques*, ser. Cambridge Studies in Speech Science and Communication, W. J. Hardcastle and N. Hewlett, Eds., Cambridge University Press, Dec. 1999, pp. 7–30. DOI: 10.1017/CBO9780511486395.002.

[50] C. A. Fowler and E. Saltzman, "Coordination and coarticulation in speech production," *Language and Speech*, vol. 36, no. 2-3, pp. 171–195, 1993. DOI: 10.1177/002383099303600304.

[51] D. Hemmerling and M. Wojcik-Pedziwiatr, "Prediction and estimation of parkinson's disease severity based on voice signal," *Journal of Voice*, 2020, in press, ISSN: 0892-1997. DOI: 10.1016/j.jvoice.2020.06.004.

[52] P. Gómez-Vilda, J. Mekyska, J. M. Ferrández, D. Palacios-Alonso, A. Gómez-Rodellar, V. Rodellar-Biarge, Z. Galaz, Z. Smekal, I. Eliasova, M. Kostalova, *et al.*, "Parkinson disease detection from speech articulation neuromechanics," *Frontiers in Neuroinformatics*, vol. 11, no. 56, pp. 1–17, 2017. DOI: 10.3389/fninf.2017.00056.

[53] B. Mirheidari, D. Blackburn, T. Walker, M. Reuber, and H. Christensen, "Dementia detection using automatic analysis of conversations," *Computer Speech & Language*, vol. 53, pp. 65–79, 2019, ISSN: 0885-2308. DOI: 10.1016/j.csl.2018.07.006.

[54] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann, "Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features," *Computer Speech & Language*, vol. 53, pp. 181–197, 2019, ISSN: 0885-2308. DOI: 10.1016/j.csl.2018.07.007.

[55] L. Brabenec, J. Mekyska, Z. Galaz, and I. Rektorova, "Speech disorders in parkinson's disease: Early diagnostics and effects of medication and brain stimulation," *Journal of neural transmission*, vol. 124, no. 3, pp. 303–334, 2017. DOI: 10.1007/s00702-017-1676-0.

[56] E. Maor, J. D. Sara, D. M. Orbelo, L. O. Lerman, Y. Levanon, and A. Lerman, "Voice signal characteristics are independently associated with coronary artery disease," *Mayo Clinic Proceedings*, vol. 93, no. 7, pp. 840–847, 2018, ISSN: 0025-6196. DOI: 10.1016/j.mayocp.2017.12.025.

[57] T. Ono, K. Hori, and T. Nokubi, "Pattern of tongue pressure on hard palate during swallowing," *Dysphagia*, vol. 19, no. 4, pp. 259–264, 2004. DOI: 10.1007/s00455-004-0010-9.

[58] M. R. Kapsner-Smith, E. J. Hunter, K. Kirkham, K. Cox, and I. R. Titze, "A randomized controlled trial of two semi-occluded vocal tract voice therapy protocols," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 3, pp. 535–549, 2015. DOI: 10.1044/2015_JSLHR-S-13-0231.

[59] E. Z. Murano, H. Shinagawa, J. Zhuo, R. P. Gullapalli, R. A. Ord, J. L. Prince, and M. Stone, "Application of diffusion tensor imaging after glossectomy," *Otolaryngology—Head and Neck Surgery*, vol. 143, no. 2, pp. 304–306, 2010. DOI: 10.1016/j.otohns.2010.03.012.

[60] D. Aalto, O. Aaltonen, R.-P. Happonen, P. Jääsaari, A. Kivelä, J. Kuortti, J.-M. Luukinen, J. Malinen, T. Murtola, R. Parkkola, J. Saunavaara, T. Soukka, and M. Vainio, "Large scale data acquisition of simultaneous mri and speech," *Applied Acoustics*, vol. 83, pp. 64–75, 2014, ISSN: 0003-682X. DOI: 10.1016/j.apacoust.2014.03.003.

[61] J. Takatsu, N. Hanai, H. Suzuki, M. Yoshida, Y. Tanaka, S. Tanaka, Y. Hasegawa, and M. Yamamoto, "Phonologic and acoustic analysis of speech following glossectomy and the effect of rehabilitation on speech outcomes," *Journal of Oral and Maxillofacial Surgery*, vol. 75, no. 7, pp. 1530–1541, 2017. DOI: 10.1016/j.joms.2016.12.004.

[62] R. I. Damper, "Speech technology—implications for biomedical engineering," *Journal of Medical Engineering & Technology*, vol. 6, no. 4, pp. 135–149, 1982. DOI: 10.3109/03091908209041006.

[63] B. Yuksekkaya, A. A. Kayalar, M. B. Tosun, M. K. Ozcan, and A. Z. Alkar, "A gsm, internet and speech controlled wireless interactive home automation system," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 3, pp. 837–843, 2006. DOI: 10.1109/TCE.2006.1706478.

[64] U. Qidwai and M. Shakir, "Ubiquitous arabic voice control device to assist people with disabilities," in *2012 4th International Conference on Intelligent and Advanced Systems (ICIAS2012)*, vol. 1, 2012, pp. 333–338. DOI: 10.1109/ICIAS.2012.6306213.

[65] D. R. Bolla, Shivashankar, T. S. Pavan, N. M. Ashwini, V. Kavya, and K. M. Mahesh, "Voice enabled gadget assistance system for physically challenged and old age people," in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, 2017, pp. 2081–2085. DOI: 10.1109/RTEICT.2017.8256966.

[66] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, E. Saltzman, and M. Tiede, "Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition," *Speech Communication*, vol. 89, pp. 103–112, 2017, ISSN: 0167-6393. DOI: 10.1016/j.specom.2017.03.003.

[67] F. H. Guenther, "A neural network model of speech acquisition and motor equivalent speech production," *Biological cybernetics*, vol. 72, no. 1, pp. 43–53, 1994. DOI: 10.1007/BF00206237.

[68] F. H. Guenther, *Neural control of speech*. USA: MIT Press, 2016.

[69] E. Saltzman and J. Kelso, "Skilled actions: A task-dynamic approach," *Psychological review*, vol. 94, no. 1, pp. 84–106, 1987. DOI: 10.1037/0033-295X.94.1.84.

[70] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, no. 4, pp. 333–382, 1989. DOI: 10.1207/s15326969eco0104_2.

[71] J. Houde and S. Nagarajan, "Speech production as state feedback control," *Frontiers in Human Neuroscience*, vol. 5, no. 82, pp. 1–14, 2011, ISSN: 1662-5161. DOI: 10.3389/fnhum.2011.00082.

[72] Audacity Team, *Audacity®: Free audio editor and recorder [computer application]*, version 2.2.2, Feb. 2018. [Online]. Available: https://audacityteam.org/.

[73] M. R. Hasan and M. M. Hasan, "Isolated bengali vowel and word speech sounds," *Mendeley Data,* V1, 2021, [Dataset]. DOI: 10.17632/2h6975kdsx.1.

[74] A. M. Selvan and R. Rajesh, "Word classification using neural network," in *Advances in Computing and Communications*, A. Abraham, J. L. Mauri, J. F. Buford, J. Suzuki, and S. M. Thampi, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 497–502, ISBN: 978-3-642-22720-2. DOI: 10.1007/978-3-642-22720-2_52.

[75] R. B. M. Baquirin and P. L. Fernandez, "Artificial neural network (ann) in a small dataset to determine neutrality in the pronunciation of english as a foreign language in filipino call center agents: Neutrality classification of filipino call center agent's pronunciation," *Inteligencia Artificial*, vol. 21, no. 62, pp. 134–144, Nov. 2018. DOI: 10.4114/intartif.vol21iss62pp134-144.

[76] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9, pp. 341–345, Jan. 2001.

[77] B. McFee, V. Lostanlen, M. McVicar, A. Metsai, S. Balke, C. Thomé, C. Raffel, A. Malek, D. Lee, F. Zalkow, K. Lee, O. Nieto, J. Mason, D. Ellis, R. Yamamoto, S. Seyfarth, E. Battenberg, Морозов, B., R. Bittner, K. Choi, J. Moore, Z. Wei, S. Hidaka, nullmightybofo, P. Friesch, F.-R. Stöter, D. Hereñú, T. Kim, M. Vollrath, and A. Weiss, *Librosa/librosa: 0.7.2*, version 0.7.2, Jan. 2020. DOI: 10.5281/zenodo.3606573.

[78] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, G. Izacard, A. Joulin, G. Synnaeve, J. Verbeek, and H. Jégou, *Resmlp: Feedforward networks for image classification with data-efficient training*, 2021. arXiv: 2105.03404 [cs.CV].

[79] J. Grus, *Data science from scratch: first principles with python*, 1st. USA: O'Reilly Media, 2019, ISBN: 978-1-49-190142-7.

[80] F. Chollet, *Deep learning with Python*, 1st. Manning New York, 2018, ISBN: 978-1-61-729443-3.

[81] S. Haykin, *Neural Networks: A Comprehensive Foundation (3rd Edition)*. USA: Prentice-Hall, Inc., 2007, ISBN: 978-0-13-147139-9.

[82] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2014. arXiv: 1412.6980 [cs.LG].

[83] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. USA: MIT Press, 2016, ISBN: 978-0-26-203561-3. [Online]. Available: https://www.deeplearningbook.org.

[84] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06, New York, NY, USA: Association for Computing Machinery, 2006, pp. 233–240, ISBN: 1595933832. DOI: 10.1145/1143844.1143874.

[85] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960. DOI: 10.1177/001316446002000104.

[86] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977, ISSN: 0006341X, 15410420. [Online]. Available: http://www.jstor.org/stable/2529310.

[87] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of Database Systems*, L. LIU and M. T. ÖZSU, Eds. Boston, MA: Springer US, 2009, pp. 532–538, ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_565.

[88] S. R. Mitra and M. M. Hasan, "Comparison of vocal-tract dynamics for bangla vowel and vowel-consonant-vowel sequence," in *International Conference on Advanced Information and Communication Technology*, 2016, pp. 1–7.

[89] M. R. Hasan and M. M. Hasan, "Investigation of the effect of mfcc variation on the convolutional neural network-based speech classification," in *2020 IEEE Region 10 Symposium (TENSYMP)*, IEEE, Jun. 2020, pp. 1408–1411. DOI: 10.1109/TENSYMP50017.2020.9230697.

Publications from this Thesis

[1] M. R. Hasan, M. M. Hasan, and M. Z. Hossain, "How many mel-frequency cepstral coefficients to be utilized in speech recognition? a study with the bengali language," *The Journal of Engineering,* IET, pp. 1–11, Oct. 2021. DOI: 10.1049/tje2.12082.

[2] M. R. Hasan, M. M. Hasan, and M. Z. Hossain, "Effect of vocal tract dynamics on neural network-based speech recognition: A bengali language-based study," *Expert Systems,* Wiley, Sep. 2021, Under Review.

[3] M. R. Hasan, M. M. Hasan, and M. Z. Hossain, "Tuning deep neural network hyperparameters for bengali speech token classification," in *International Conference on 4th Industrial Revolution and Beyond (IC4IR 2021)*, Under Review, Oct. 2021.