ARTICULATION MODELLING OF VOWELS IN DYSARTHRIC AND NON-DYSARTHRIC SPEECH

by

Rahaf Albalkhi

B.Sc. (Hons), Arab Open University, Amman, 2017

A Thesis Submitted in Partial Fulfillment of the Requirements
of the Degree of

MASTER OF APPLIED SCIENCE
in the Department of Electrical and Computer Engineering

ARTICULATION MODELLING OF VOWELS IN DYSARTHRIC AND NON-DYSARTHRIC SPEECH

by

Rahaf Albalkhi
B.Sc. (Hons), Arab Open University, Amman, 2017

Supervisory Committee

Dr. Mihai Sima, Supervisor
Department of Electrical and Computer Engineering

Dr. Nigel Livingston, Supervisor
School of Public Health and Social Policy

ABSTRACT

People with motor function disorders that cause dysarthric speech find difficulty using state-of-the-art automatic speech recognition (ASR) systems. These systems are developed based on non-dysarthric speech models, which explains the poor performance when used by individuals with dysarthria. Thus, a solution is needed to compensate for the poor performance of these systems. This thesis examines the possibility of quantifying vowels of dysarthric and non-dysarthric speech into codewords regardless of inter-speaker variability and possible to be implemented on limited-processing-capability machines. I show that it is possible to model all possible vowels and vowel-like sounds that a North American speaker can produce if the frequencies of the first and second formants are used to encode these sounds. The proposed solution is aligned with the use of neural networks and hidden Markov models to build an acoustic model in conventional ASR systems. A secondary finding of this study includes the feasibility of reducing the set of ten most common vowels in North American English to eight vowels only.

Contents

List of Tables

List of Figures

Acknowledgments

I acknowledge with respect the Lekwungen peoples on whose traditional territory the university of Victoria, where I was able to conduct this research, stands and the Songhees, Esquimalt and WSÁNEĆ peoples whose historical relationships with the land continue to this day.

I started this research never knowing that one day I would benefit directly from its results. For that, I am grateful to the exceptional people who supported me throughout this journey.

I thank my supervisors Dr. Mihai Sima and Dr. Nigel Livingston for their enormous support when I needed it, encouragement, and guidance throughout the journey. And I acknowledge the rich input from my colleagues at the assistive technology laboratory. I also thank Dr. Nikitas Dimopoulos for the rich discussions.

This work was funded in part by Mathematics of Information Technology and Complex Systems (MITACS).

I thank Karen Short and Tim Smith for the extraordinary guidance and support, and for being exceptional people. And I thank Lorna Sandler, and Derek Doyle for all the rich conversations. I am grateful to have them in my life.

I thank the World University Services of Canada (WUSC) for their unique mission, Dr. Marlea Clarke and Dr. Scott Watson, and the UVic WUSC local committee.

This thesis would have been impossible without the support of the volunteers who generously donated their voice samples.

I acknowledge with gratitude the conversations with Corina Botelho and Thiago H., that introduced me to linguistics.

The people I have met in Victoria played a key role shaping this work. I thank Max, Doug, and the Anderson family for the care and support. The Hajibrahim family, Noelle Mason, Yeshua Moser-Puangsuwan, and Anitra Puangsuwan-Moser for being great friends. And the amazing Sophia Papp for sharing the experience.

Most importantly, I would like to thank my mother whose belief in this research fueled the journey. And I thank my father and my sisters Reem and Shaimaa for all the support and encouragement, and my sister Sara, for willing to sacrifice her own success for the sake of my wellbeing.

Dedication

To my brother, Mouaz Albalkhi, my inspiration

Chapter 1

Introduction

Speech commands are one way to interact with the machines around us to perform certain tasks. Some people consider using their voice to complete a task is fancy, while others consider it a necessity to improve their quality of life and increase their independence. Thus, developing a robust Automatic Speech Recognition (ASR) system that works for different speech types is a necessity. Dysarthric speech results from motor function disorders that cause a disturbance in the control over speech articulators [1]. Although dysarthric speech is linguistically normal, the articulation of speech is unclear [2]. Therefore, human listeners and machines may find it difficult to recognize dysarthric speech. Most state-of-the-art ASR systems are built based on models of non-dysarthric speech. This explains the poor performance of these systems when used by individuals with dysarthria. In some cases, these systems are completely incapable of recognizing even a simple isolated word.

A popular approach to solve this problem, is to use Artificial Neural Networks (ANNs). Researchers working on tasks mimicking the recognition abilities of humans, such as speech recognition and image recognition, commonly use ANNs to solve these problems. The ability of the ANNs to learn and predict or classify patterns makes them a powerful solution to many problems. Despite the powerfulness of ANNs, Researchers could not, yet, adjust ANNs based models to recognize all degrees of dysarthric speech using a single model. This raises the question whether or not a robust model of speech data that accommodates all degrees of dysarthric speech

from severe dysarthria to no dysarthria in a certain language regardless of the inter-speaker variability and without the need for strong processing capabilities and a large amount of data exists.

To explore answers to this question, the characteristics of dysarthric speech need to be understood. Unclear articulation is the main identifier of dysarthric speech, which makes the problem of dysarthric speech recognition an articulation problem. Thus, the incorporation of articulatory knowledge in an ASR system is necessary to reflect the different articulation in the spectrum of dysarthric speech, which facilitates solving the problem. Conventionally, ANNs are fed with features based on speech perception mechanisms, such as the Mel Frequency Cepstral Coefficients (MFCC), which does not reflect the differences in the dysarthric speech spectrum. A direct way to obtain articulatory knowledge is to use electromagnetic articulography (EMA) to capture the state of the articulators while producing sounds. Nevertheless, this is an invasive technique. Another way to obtain articulatory knowledge is by using a microphone to collect speech signals and then perform an acoustic-to-articulatory inversion process.

A characteristic of dysarthric speech is that individuals with severe dysarthria can produce vowels easier than any other phoneme. Some phonemes are produced with a greater control over the articulators. For instance, phonemes that require full closure of the mouth like the consonant /P/ or phonemes that are produced by nasal coupling like the nasal /M/. These phonemes can be the most unintelligible sounds in dysarthric speech and can be the hardest to be recognized by a machine. In general, vowel production requires the least control degree over the articulators compared to other phonemes. This problem can be approached by developing a vowel-oriented ASR system. This is possible since vowels hold information about the production process of the vowel and its neighboring phonemes. This phenomenon is called coarticulation and can be used in

the recognition of dysarthric speech. Once vowels are correctly recognized, the task of identifying other phonemes is simplified. Thus, the focus of the proposed solution in this research is on vowels.

1.1. Approach

The acoustic model of a conventional ASR system is trained with a large amount of data based on the speech perception mechanism. The basic architecture of an ASR system is illustrated in Figure 1.1 below. The proposed solution replaces the conventional acoustic model with a model that has only two parameters and is based on the sound articulation mechanism.

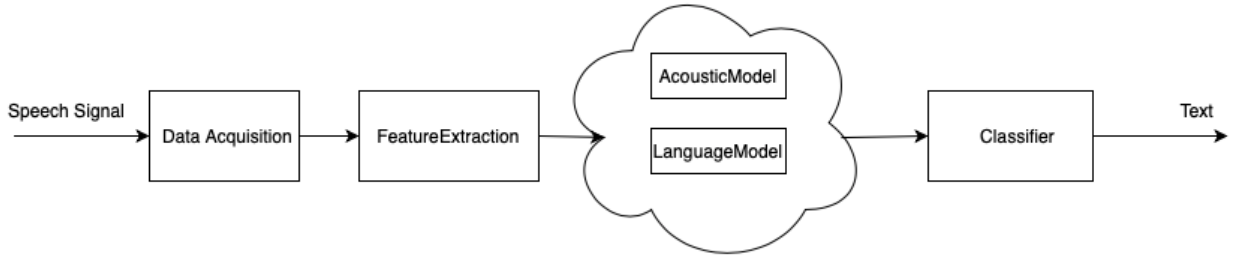


Figure1.1: The basic architecture of typical ASR systems.

Figure 1.1 depicts the architecture of conventional ASR systems. First, the speech signal is acquired through a microphone. Then, features are extracted from the signal. The extracted features are then processed according to the acoustic and language models. The acoustic model is responsible for interpreting the features into phonemes and the language model is responsible for identifying all possible sequences of words. Both models are usually based on ANNs, and the acoustic model is usually built to mimic speech perception which is non-adjustable to accommodate all degrees of dysarthric speech. The outcome of these models is passed through a classifier that outputs text.

The features extracted from an acoustic signal hold information about the vocal tract's configuration state. It will be shown that only two features are enough to model the vowels and

vowel-like sounds that are possible to be produced by all possible combinations of jaw and tongue positions. The jaw opening is quantified by the frequency of the first formant (F1), and tongue position is quantified by the frequency of the second formant (F2). These two frequency-domain features are used to quantify all possible vowel and vowel-like sounds produced by any North American English speaker into codewords, and to build a vowel-oriented acoustic model. This model uses the minimum number of features and requires minimum processing capabilities to allow operation on machines as small as a smart watch.

1.2. Thesis Outline

Some of the thesis chapters are written as papers. Namely Chapter 4, Chapter 5, and Chapter 6. Other chapters are written as normal thesis chapters. The thesis is organized as follows: Chapter 2 provides a background of the concepts used in the thesis. Then, Chapter 3 lists the related works in the area of speech recognition in general and the area of dysarthric speech recognition in particular, emphasizing works that incorporated articulatory knowledge. Chapter 4 explores the recognition of vowels based on the relation between jaw opening width and frequency of the first formant, as determined from the results in [3]. The set of ten common vowels in North American English is divided into groups based on the mean F1 frequency of each vowel. Similarly, in Chapter 5, the set of ten common vowels in North American English is divided into groups based on the mean F2 frequency of each vowel. F2 is considered to quantify the tongue position based on results from [3]. After the study of features in Chapters 4 and 5, Chapter 6 discusses how F1 and F2 are used to provide a domain of codewords, each representing a vowel or a vowel-like sound. The domain is used to recognize vowels in North American English regardless of inter-speaker variability and may be implemented on limited-processing-capability machines without the need

for a large amount of data to build the domain. This chapter also explains how this domain is used to recognize the rest of the phonemes in North American English. Finally, Chapter 7 provides a summary of the contributions that resulted from this research and proposes how the findings from Chapter 6 may be used in the future in commercial or clinical applications. In addition, Chapter 7 discusses the areas of this research where further investigation might be conducted.

Chapter 2

Background

This chapter presents a more thorough view of the fundamental concepts used in this thesis. The first section studies the sound production process as a physical movement of speech organs and describes the nature of these sounds. It also clarifies when a sound is considered a language unit. Section 2.2 discusses the way humans and machines process the received sound signals as speech. Section 2.3 presents information about dysarthric speech in terms of its nature and how it differs from non-dysarthric speech.

Terminology: The term "sound production" is used throughout this thesis instead of "speech production" when talking about the production process only. The reason for using "sound" instead of "speech" is that speech is a process that includes the planning stage in the brain, while sounds simply arise from air flowing through the physical configuration of the vocal system. In developing ASR systems, the planning stage in the brain is not included. Thus, if we refer to the sound as a result of the articulators' movement, it is more accurate to use the term "sound" instead of "speech". Once the listener perceives the sound and process it to retrieve speech, we can use the term "speech" because the processing stage in the brain is involved. For example, Section 2.2 uses the term "Speech Processing" as it handles the processing of sounds in the brain to retrieve speech.

2.1. Sound Production

Sound is generated by air flowing through the vocal system. The production of different sounds depends on different vocal system shapes and configurations as in musical instruments. The energy of the sound is controlled by the abdominal muscle that controls the amount of air entering the lungs [4]. The shape of the vocal system is determined by the shape of its parts; the vocal tract, and the nasal tract. The vocal tract starts at the glottis and ends at the lips, containing the pharynx and oral cavity. The vocal tract length is typically 17 cm in an adult male, and its cross-sectional area is around the range of 0-20 cm^2 [4]. The nasal tract starts at the velum and ends at the nostrils and is about 12 cm long [4]. Lowering the velum couples the two tracts, which is a technique used to produce nasals, a class of sounds that will be discussed later. There are more than 100 muscles that can be active simultaneously during one second of speech, which makes no sound produced the same each time it is uttered [5]. Figure 2.1 below shows the main articulators of the vocal system and sub-system that are responsible for sound production. In this thesis, the effect of two main articulators is examined. The two articulators, namely the jaw and the tongue, contribute to the sound properties by controlling the vocal tract cross sectional area. The jaw muscle can move in a complex three-dimensional manner, controlled by the masseter, temporalis, and medial pterygoid muscles [6]. Indirectly, the jaw assists the positioning of the tongue and lips [42]. The tongue is the most complex articulator consisting of 12 muscle pairs, it is flexible, and it takes about 50 ms to change from one position to the another [7]. The following subsections discuss two important aspects of the sounds: the classification based on the mechanism to produce the sounds, and how they can be considered part of a language.

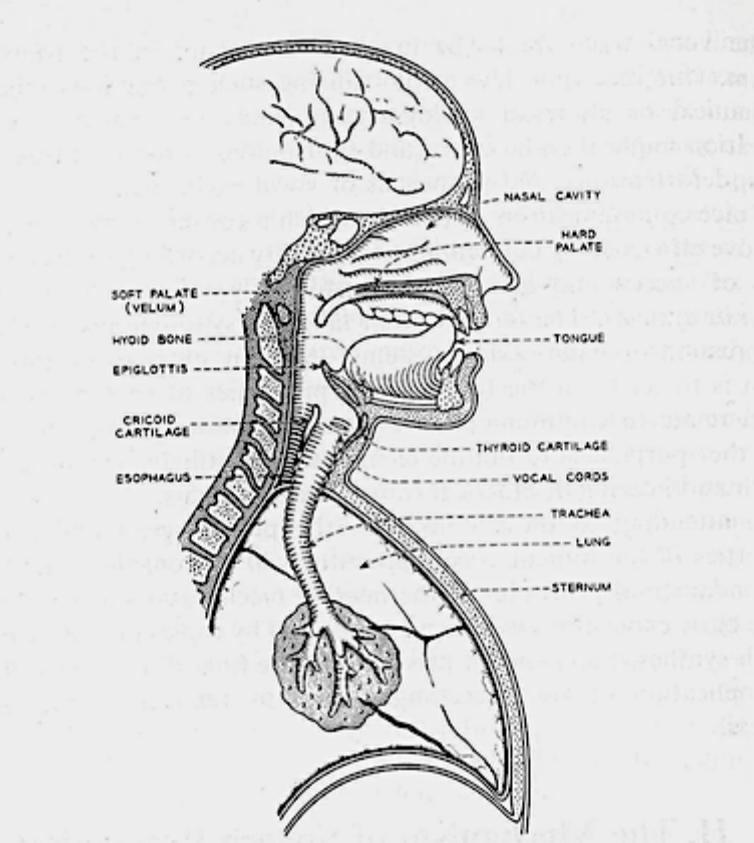


Figure 2.1: The articulators involved in the sound production process (From [4]).

2.1.1.Classification of Sounds Based on their Production Mechanism

The books written by Rabiner and Schaffer [8], and Flanagan [4] provide information on the mechanism of sound production and form the basis of this subsection. Sounds can be classified according to the mechanism of producing them. For example, in [8] it is suggested that sounds can be classified according to their *excitation mode* into three classes. The first class is *Voiced* sounds, which are identified by the vibration of vocal cords (the cords that define the glottis opening). The properties of voiced sounds are dependent on the period of the oscillation, which is determined by the mass and compliance of the cords and the subglottal pressure. A common voiced sound property is pitch, which is defined as the frequency of vocal cords vibration. The second class is

Unvoiced sounds, which do not require vocal cords vibration. Finally, the third class is *plosives* where pressure builds up behind a certain complete closure point, then is abruptly released. Plosives have distinct waveforms as they have a small amplitude and are preceded by a gap of no amplitude that represents the time of complete closure of the vocal tract [8]. Figure 2.2 below shows a plosive.

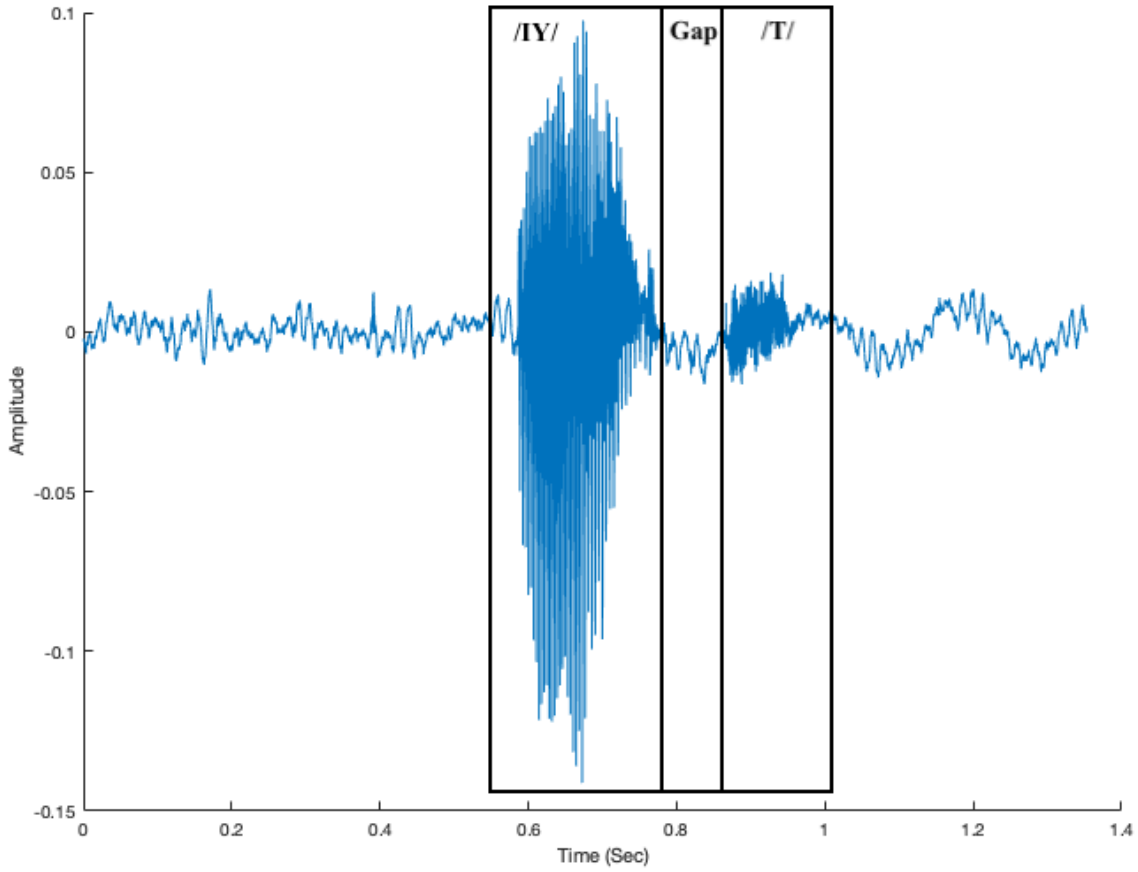


Figure 2.2: Waveform of the word "Eat" uttered by speaker #4.

Figure 2.2 corresponds to data collected from speaker #4, a female who speaks English as a first language and have no dysarthria. There are two sounds in Figure 2.2. The first sound is /IY/, which is a voiced sound. The second sound is /T/, which is a plosive. The first part which is a voiced sound shows a large amplitude, followed by a gap of no amplitude which represents full

closure of the vocal tract. The gap suggests that the sound to come is a plosive. The plosive sound /T/ has a smaller amplitude than the voiced sound /IY/.

An alternative way to classify sounds is by examining the *continuity of the tract configuration***.** There are continuant and non-continuant sounds according to the vocal tract configuration. To produce *continuant* sounds, the vocal tract shape stays the same for the whole sound production duration. On the other hand, *non-continuant* sounds have a changing vocal tract shape. For example, diphthongs are non-continuant sounds, and the production process starts by having a shape of a vowel and ends by having a shape of another vowel (more information about diphthongs will be provided later).

The most common classification method is based on *the manner and place of the articulators***.** Sounds are divided into four main classes: vowels, diphthongs, semivowels, and consonants. The consonants class is further divided into five subclasses. *Vowels* result from quasi-periodic pulses of air caused by vocal cords vibrations [8] travelling through the oral cavity, they are voiced and continuant sounds. Conventionally, the acoustic signals of vowels are distinguished based on the tongue hump position and tongue hump height. Table 2.1 below provides information about the tongue state for ten vowels.

VOWELS			
Degree of tongue constriction (Tongue hump height)	**Tongue hump position**		
	Front	**Central**	**Back**
High	/IY/ as in "Eat"	/ER/ as in "Bird"	/UW/ as in "Two"
	/IH/ as in "It"		/UH/ as in "Hood"
Medium	/EH/ as in "Ed"	/AH/ as in "Up"	/AO/ as in "Ought"
Low	/AE/ as in "At"		/AA/ as in "Odd"

Table 2.1: Tongue state for the set of ten vowels in North American English (After [4]).

The first column in Table 2.1 provides three possible states of tongue hump height ("High", "Medium", or "Low"), while the first row provides three possible states of tongue hump position in the oral cavity. "Front" which means close to the mouth opening, "Central", or "Back" which means closer to the velum. The elements of the table are read according to the column and row. For example, the sound /IH/ has a high tongue hump, that is positioned in the front of the mouth. Table 2.1 also shows that vowels /UW/ and /UH/ cannot be uniquely identified based on information about tongue hump height and tongue hump position. The vowels /IY/ and /IH/ cannot be uniquely identified either.

Diphthongs are produced in the same way as vowels since they are a combination of two vowels. For instance, the diphthong /OY/ as in "Toy" starts with the shape of the vowel /AO/ and ends with the shape of the vowel /IY/. The place of articulation of a diphthong is the same place

of articulation as the two vowels of which it is comprised. There are approximately five diphthongs in North American English.

Semivowels are produced by a gliding transition. There are four semivowels: /R/ as in "Read", /W/ as in "We", /Y/ as in "Yield", and /L/ as in "Lee".

The last class is *consonants,* which is further divided into five subclasses: nasals, fricatives, stops, whisper, and affricates. Nasals are voiced sounds that propagate for most of the time through the nasal tract rather than the vocal tract. This is made possible by lowering the velum and connecting the vocal and nasal tracts together. The place of articulation for each of the three nasals is shown in Table 2.2 below.

NASALS	
Place of articulation	**Sound**
Labial	/M/ as in "Me"
Alveolar: the tongue at the gum ridge	/N/ as in "Knee"
Palatal/Velar: tongue against hard palate	/NG/ as in "Ping"

Table 2.2: Nasals and their place of articulation (After [4]).

Table 2.2 shows the articulators involved in the production of nasals and the state of these articulators. The first column gives the place of articulation, and the second column contains the list of nasalized sounds. The velum, the lips, and the tongue are the main articulators that contribute to the production of a nasal.

Fricatives are produced by a turbulent air flow at different constriction points as shown in Table 2.3. Fricatives can be either voiced or unvoiced.

FRICATIVES		
Place of articulation	**Voiced**	**Unvoiced**
Labio-dental: upper teeth on the lower lip	/V/ as in “Vee”	/F/ as in “Fee”
Dental: tongue behind the teeth	/DH/ as in “Thee”	/TH/ as in “Theta”
Alveolar: the tongue at the gum ridge	/Z/ as in “Zee”	/S/ as in “Sea”
Palatal: tongue against hard palate	/ZH/ as in “Vision”	/SH/ as in “She”

Table 2.3: Fricatives and their place of articulation (After [4]).

Table 2.3 describes the state of the involved articulators of eight different fricatives. The first column indicates the place of articulation of these sounds, while the second and third columns categorize the fricative sounds as voiced or unvoiced respectively.

Lastly, Stops are the result of a sudden release of trapped air behind a certain point of constriction in the oral cavity, they can be either voiced or un-voiced. Affricates are a combination of two other phonemes. The whisper sound /HH/ is a noise-like sound that is highly influenced by the adjacent sound. The place of articulation for stops and the whisper sound /HH/ are shown in Table 2.4. Because affricates are a combination of other phonemes, the place of articulation of an affricate can be derived from the place of articulation of the phonemes of which it is comprised.

STOPS AND WHISPER		
Place of articulation	**Voiced**	**Unvoiced**
Labial	/B/ as in “Bee”	/P/ as in “Pen”
Alveolar	/D/ as in “Dee”	/T/ as in “Tea”
Palatal/Velar	/G/ as in “Green”	/K/ as in “Key”
Glottal: vocal cords constricted and fixed		/HH/ as in “He”

Table 2.4: Place of articulation for stops and whisper sounds (After [4]).

The first column in Table 2.4 provides the place of articulation for stops and the whisper sound. The second and third columns classify these sounds as voiced or unvoiced, respectively.

2.1.2. Phonemes as Language Sounds

The sounds produced based on the mechanism discussed earlier in Section 2.1.1 are not the only sounds that humans can produce. In fact, the previous section only discussed the mechanism to produce sounds that are meaningful in non-dysarthric North American English. Different languages have different mechanisms and usage of articulators. If a certain sound distinguishes a word from another in a certain language, it is called a “phoneme”. A phoneme is “any of the abstract units of the phonetic system of a language that correspond to a set of similar speech sounds (such as the velar \k\ of cool and the palatal \k\ of keel), which are perceived to be a single distinctive sound in the language” [9]. Phonemes can also be seen as a unique code for the articulatory gestures which depends on the language and dialect [4]. Listeners perceive a phoneme as the basic unit of *speech* that is processed by the brain. In a language, phonemes are used to build

larger blocks called syllables that are then used to form a word. A phoneme has different possible pronunciations called “allophones” and would be called a “phone” if the language specifications are disregard.

Historically, linguists have studied the characteristics of these phonemes for the purpose of proper classification, and this has raised the need to have an alphabet that contains all possible phonemes in a certain language. The English language has no fixed number of phonemes. Nevertheless, most textbooks claim that there are 44 phonemes [10], while in [8], it is suggested that there are 42 phonemes in North American English. The International Phonetic Alphabet (IPA) chart (See appendix A) provides a standard notation based on the Latin alphabet for possible phonemes in different languages and is used by linguists to describe a language. However, the IPA chart is not computer-friendly, which means there is a need for an alphabet that uses ASCII characters. The ARPABET phonemes set (See appendix B) was developed to address this issue. The phonemes are described using ASCII characters. The notation from ARPABET is used throughout this study.

The focus of this study is on vowels. Since there is no agreed number of vowels, the classification of three resources is compared to produce the list of vowels of interest for the purpose of this research. The first resource is the CMU Pronouncing Dictionary phoneme set (See appendix C for the full list of phonemes in the CMU set). It is a phoneme set based on the ARPABET, which contains 39 phonemes and it is used in some automatic speech recognition applications. The second resource is the vowel classification in [8], a book on the digital processing of speech signals. The third resource is the classification in [4], a book on speech analysis, synthesis and perception. Table 2.5 below shows 15 phonemes where at least one of the three resources has classified as vowels.

Row	The Phoneme	Class in CMU Set [11]	Class in [8]	Class in [4]
1	/IY/	Vowel	Vowel	Vowel
2	/IH/	Vowel	Vowel	Vowel
3	/EH/	Vowel	-	Vowel
4	/AE/	Vowel	Vowel	Vowel
5	/AA/	Vowel	Vowel	Vowel
6	/AH/	Vowel	Vowel	Vowel
7	/AO/	Vowel	Vowel	Vowel
8	/ER/	Vowel	Vowel	Vowel
9	/UW/	Vowel	Vowel	Vowel
10	/UH/	Vowel	Vowel	Vowel
11	/AY/	Vowel	Diphthong	Diphthong
12	/OY/	Vowel	Diphthong	Diphthong
13	/AW/	Vowel	Diphthong	Diphthong
14	/EY/	Vowel	Vowel	Usually Diphthong
15	/OW/	Vowel	Vowel	Usually Diphthong

Table 2.5: The vowels according to [11], [8], and [4].

The first column in Table 2.5 provide the row number. The second column provides a list of phonemes that are considered vowels in at least one of the resources. The third, fourth and fifth columns show the classification of these phonemes according to [11], [8], and [4], respectively. By comparing the classifications of vowels in the three sets in Table 2.5, it is found that some

vowels in the CMU set are classified as diphthongs in [8] and [4]. Since diphthongs can be seen as a combination of two vowels, the vowels that are classified as diphthongs in at least one of the three resources are eliminated. Thus, vowels in rows 1 through 10 in Table 2.5 are used as the vowel set that will be studied in this thesis.

2.2. Speech Processing

In a conversation between two humans, one speaks while the other listens. The speaker produces a signal that is a string of sounds using the mechanism discussed earlier in Section 2.1.1 and the listener processes the received string of sounds using their brain. When the listener is a machine, the machine processes the received string of sounds using techniques that mimic a human listener's auditory and perceptual systems. In both cases, if the string of sounds forms a meaningful word (according to the human's /machine's knowledge) it is considered speech. Otherwise, it is simply a sound. In the following section, *speech* processing is discussed for two different listener types, a human listener, and a machine listener.

2.2.1. Speech Perception by Humans

Although it is not clear how humans exactly process information that leads to understanding the pattern of a certain sound [12], some interesting findings can be used. The process of understanding aural patterns in the human brain is considered to be a psychophysical problem because humans try to associate the perceived sound signal with past experience [13]. This process of recognizing patterns not only relies on previous experiences, but also on the feature extraction, contextual understanding, and memory [13]. The process of extracting the discriminatory features from a string of sounds signal is naturally considered as a combination of

structural analysis, feature tuning that emphasizes some important structural properties, and probability optimization strategies [14]. This implies that the decision rules are also part of the feature extraction process [14]. The classification process is then based on a number of dimensions, such as loudness, timbre and pitch [12]. Three main assumptions are made with regards to how our brains process the speech signals. First, it is not a single process, but a combination of multiple processes over time that begins with the reception of the signal and ends at the response of the receiver. Second, the human brain has a capacity and processing limitations when it comes to manipulating sensory data. Third, memory processes are involved in speech perception at different stages of the perception [15].

Speech is assumed to be perceived by the listener as a signal that contains clues about the production process. Such clues assist in future processing steps even if there is a mistake while producing the sound, i.e., if a person mistakenly replaced a phoneme by another, it is still understood [16]. On the other hand, visual observation of the articulators while producing a phoneme can provide a clue about the next phoneme. According to Benguerel and Adelman [17], the rounding of the mouth can provide helpful information for the listener to anticipate the following phoneme in terms of coarticulation [17]. Interestingly, understanding the intended meaning of a communication process not only depends on the speech signal, but also on the facial expressions and body gestures [5].

2.2.2. Speech Signal Processing in Machines

Speech signals are processed digitally using algorithms that take into consideration the behavior of both the auditory and perceptual systems of humans (as mentioned in Section 2.2.1). These algorithms are the fundamentals of automatic speech recognition systems [18]. An example

of a technique that takes into consideration the auditory system is the Mel filter bank, a filter that emphasizes the content of lower frequencies and generalizes the content of higher frequencies, just as the cochlea does in the human ear. An example of mimicking the perceptual system is identifying the end and start of a word in a sentence. Once a string of sounds is recognized, the word ends, and a new word starts. For instance, our brains can still recognize the words within a hashtag despite the fact that they are not separated (e.g., "#educationchangestheworld"). Once a set of letters forms a meaningful word, our brains suggest the end of the word and the start of another. The first step to process the signal is to extract meaningful features, and then form a pattern that can be classified. The following subsections discuss two speech signal analysis domains: the time domain and the frequency domain.

2.2.2.1. Time-Domain Analysis

In order to process auditory signals, a number of steps must be followed to prepare the signal. First, the signal must be sampled, framed, and windowed. The steps are described in more detail in chapter 2 of [8]. In this section, a summary of the steps will be provided. This section also provides two common methods to analyze the signal in the time domain.

Sampling [8]: An analog signal (speech) can be represented using a finite number of samples that are taken periodically. The period is measured in Hz and is chosen based on the required resolution. The higher the period, the higher the resolution of the signal. Reconstruction of the analog signal from the samples is possible when the sampling rate is at least twice the Nyquist frequency. By sampling, we can represent analog signals as numbers that are taken periodically.

Framing and Windowing [8]: Signal analysis can be performed over the whole signal; however, it is more efficient if we can process smaller chunks of the signal, one at a time. The process of cutting the signal into a number of same-length chunks is called framing. The chunks are called frames, and each frame consists of a number of samples. Framing a signal without any processing will result in sharp edges, which will cause a loss of information at the edges of the frame. Thus, we need to pass the frame through a system that will reduce the amount of lost information at the framing process. In this case, these systems are called window functions. There are different kinds of windows that provide different results according to the application for which it is used. Some examples of windows are the rectangular window and the Hamming window.

Time domain processing methods involve direct analysis of the waveform of speech signals. They allow us to get information such as the excitation mode. Two of the time-domain processing techniques are short-time energy and zero crossing rate.

Short-time energy [8] amplifies the amplitude variations. Usually, voiced sounds have higher amplitude than un-voiced sounds, which makes it beneficial to use short-time energy to differentiate voiced and un-voiced sounds. Short-time energy also provides information about the exact time of voiced to un-voiced transitions in a speech signal. In the case of high-quality speech signals, it is possible to differentiate noise, speech, and silence. The nature of short-time energy depends on the window type and duration.

Short-time average zero-crossing rate [8] is a measure of the frequency content of a signal. The crossing happens when two adjacent samples have different signs. The energy of voiced speech is concentrated below about 3 kHz, and the energy for unvoiced speech is at higher frequencies. The higher the frequency, the higher the zero-crossing rate, which implies that for high zero crossing rates, the speech is considered unvoiced. However, it is insufficient to use the

crossing rate alone to decide whether or not a sound is voiced, the usage of a combination of techniques yields better results.

2.2.2.2. Frequency-Domain Analysis

Transformations [8]: Frequency domain representation of signals is an important way to analyze and design systems. Some of these representations are: the z-transform (or direct transform), the Fourier transform which can be seen as a special case of the z-transform that is restricted to the unit circle, and the discrete Fourier transform (DFT), where periodic sequences can be represented as a discrete sum of sinusoids. All signals behave as periodic when represented using DFT. Some of the important uses are in computing spectrum estimates, and in implementing digital filters.

Linear Predictive Coding (LPC) [18], [8]: Linear Prediction (LP) is a parametric modelling technique that approximates a signal as a linear combination of its past samples. The weighting coefficients that are defined for use in the linear combination are the predictor coefficients. The order of the predictor is the same as the number of these coefficients. Notably, the higher order predictor models are better fitted to the original spectrum and represent more details in the spectrum. The difference in the values between predicted model and actual value, i.e., the error in the model, is used as a quality indicator of the model, such as the lower the error, the more accurate the model. LPC is useful for many applications like speaker verification, speaker identification, speech classification, and formant analysis applications. LPC will be used in Chapter 4 and Chapter 5 to estimate the formants of vowels.

Formants: The frequency energy groups in a signal are called formants. Different vocal tract configurations result in different formant frequencies. On a frequency domain representation

of a signal, formants are the peaks such as the peak is the center of the formant. On a spectrum, formants are seen as defined intense energy regions. For vowels, we can usually identify five formants, which are the first five prominent peaks on the frequency representation. Linguists consider that only the first three formants are of interest. Different vowels have different formant patterns. Figure 2.3 shows an example of formants.

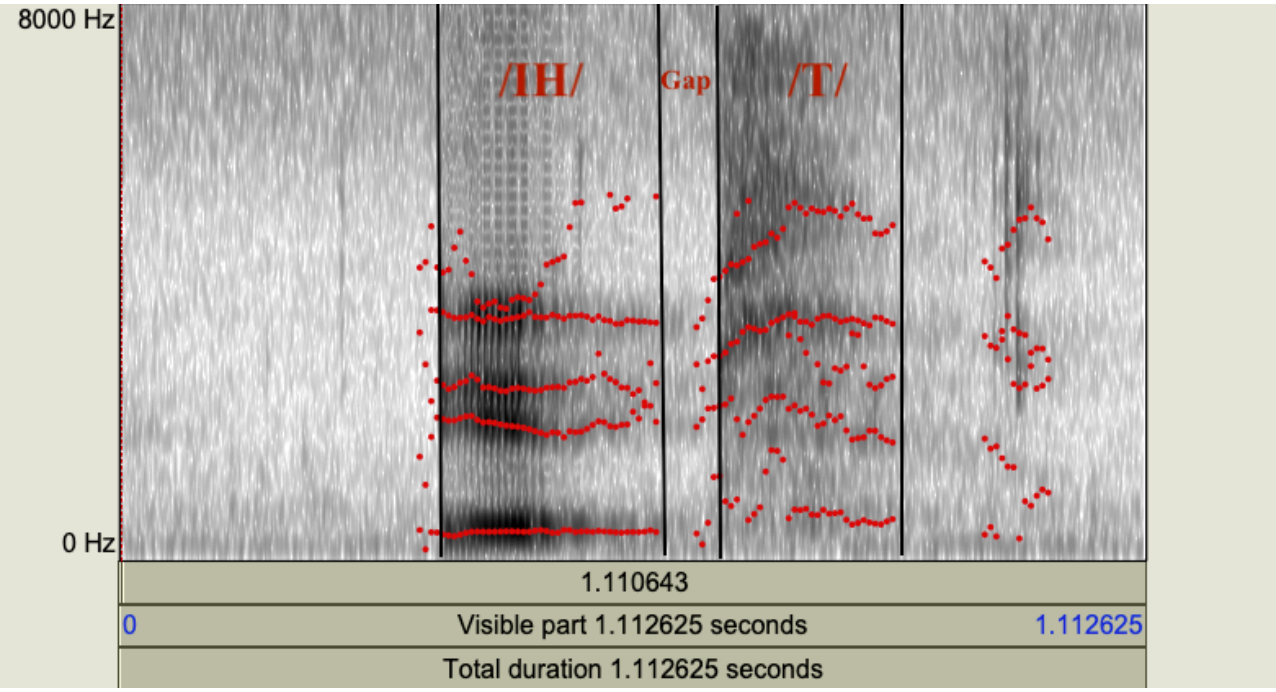


Figure 2.3: Formants of the phonemes of the uttered word "It" by speaker #7.

Figure 2.3 shows the spectrum of the word "It" uttered by Speaker #7, a male who speaks English as a first language and have non-dysarthric speech. There are two phonemes in this word: the vowel /IH/ and the plosive /T/. The red dotted lines represent the formants of the phoneme. As can be seen from the figure, the vowel /IH/ has four obvious intense regions indicated by the four dotted red lines.

2.3. Dysarthric Speech

Section 2.1 focused on the sound production process, and Section 2.2 discussed how the signal is processed in our brains or in machines. However, this applies to the case of typical sounds produced by speakers with no dysarthria, and it only applies in part to speakers with dysarthric speech. This section provides information about dysarthric speech and how the sound production process and speech processing differ from those in non-dysarthric speech.

Speech is a result of a motor function. The brain sends signals to the articulatory muscles to move in a certain way to produce a certain phoneme. When a person has a condition where the central nervous system is damaged, it is usually accompanied by motor function disorders. These disorders affect the movement of the articulatory muscles resulting in dysarthric speech. The severity of dysarthria is identified by the degree of loss of control over the articulatory muscles responsible for speech. It ranges from mild to acute. People with motor function disorders such as multiple sclerosis (MS), cerebral palsy (CP), Parkinson's disease (PD), and amyotrophic lateral sclerosis (ALS), etc., often have dysarthric speech.

Dysarthric speech is linguistically normal, but *the articulation* of speech is unclear [2]. thus, the sound production process in dysarthric speech is close to that in non-dysarthric speech, but it is not the same. For example, a speaker may not be able to produce a nasalized sound such as /M/ because of the lack of control over the velum. Another example is when a speaker intends to produce a sound that requires a full closure of the lips such as the phoneme /P/, but the range of movement of the muscles controlling the lips does not contract fully to allow for complete closure. The resulting sounds of these two examples might be unintelligible, which is dependent on whether or not the listener have previously heard this sound and known what it represents. If the listener receives a sound that has never been introduced to their brains before, or the extracted features of

that sound were not used in the design of the classifier in a machine, then it might not be possible to decode the sound as speech. A motor function disorder may also affect the flatness of produced speech, its softness, and its pitch's height or speed [19]. All of these effects are non-typical conditions of speech where the listener requires more information in order to decode the received signal as speech. Thus, when the listeners process dysarthric speech, they need to use multiple senses simultaneously to collect as much information as possible. For example, in noisy speech environments and in the case of dysarthric speech, listeners may rely on lip reading in addition to listening.

So far, no effective method has been developed to recognize dysarthric speech. The design of a system that automatically recognizes dysarthric speech must be based on traits that are mutual in both dysarthric and non-dysarthric speech. Properties like the period of a phoneme or the pitch are not standards in dysarthric speech. Thus, the system should not be designed based on these properties.

Chapter 3

Related Work

It has been almost sixty-eight years since the first speech recognition engine was developed. The engine, called "Audrey", was developed by three Bell Labs scientists in 1952 [20]. Since then, there have been significant improvements in techniques pertaining to speech recognition, allowing better performance. However, there have been no significant improvements in performance quality with regards to dysarthric speech recognition. In the first section, some of the most popular techniques that are used in state-of-the-art ASR systems are discussed briefly, while the second section focuses on previous attempts to recognize dysarthric speech. Lastly, the third section focuses on the more directly related methods to the techniques used in this research - the first one being extracting articulatory movement out of an acoustic signal, and the second one being the usage of articulatory information in the recognition of dysarthric speech.

3.1. Automatic Speech Recognition Techniques

For the past twenty years, two techniques have been the focus of ASR research: The Hidden Markov Model (HMM) and the use of ANNs. HMMs are used to build a model for speech signals; they are based on a hidden Markov process that is observed by another Markov process [21]. The acoustic models used are based on Gaussian Mixtures. Deep Neural Networks (DNNs) have shown rapid progress over the Gaussian Mixture Models due to the fact that DNNs are more powerful and have greater potential to improve the recognition quality [22]. To improve the learning of an ANN, several improvements can be done. Such as better optimization, better types of activation

functions and architecture, and more appropriate pre-processing, all of which makes DNNs more favored to recognize languages and dialects [22]. ASR is becoming more ubiquitous in the form of a virtual personal assistant, perhaps the most popular ones being Siri [23], Amazon Alexa [24], and Google Assistant [25].

3.2. Automatic Speech Recognition Techniques for Dysarthric Speech Signals

Numerous attempts have been made to improve the performance of speech recognition of dysarthric speech. This problem has been approached from many different angles, such as focusing on the speech signal itself, or on the architecture of the neural-network-based ASR systems … etc. State-of-the-art research in this field is described in this section.

Speech-to-speech approaches are used to improve the intelligibility of dysarthric speech. For example, the work in [26] focuses on modifying the dysarthric speech signal through correction of pronunciation errors, removal of repeated sounds, insertion of deleted sounds, devoicing of unvoiced phonemes, adjusting the tempo of speech, and the adjustment of frequency characteristics of speech. This latter approach modifies the speech signals itself rather than modifying speech models and classification algorithms to work for dysarthric speech. Another example of speech-to-speech systems is the work in [27]. This approach uses a network that maps the input spectrogram to another spectrogram without performing a speech-to-text step. The output is then used as an input to a speech-to-text system. This model normalizes different speech patterns into speech with a fixed accent and consistent articulation. This approach requires strong processing power.

Other attempts took the audible characteristics of speech into consideration. For instance, the work in [28] included pronunciation patterns in the recognition process by integrating confusion matrices estimates. This approach weights the response of an ASR system after setting different language model restrictions to enhance the dysarthric speech recognition. Another example to identify acoustic and articulatory distinctiveness of vowels and consonants in non-dysarthric speech is found in the study conducted in [29]. EMA was used to derive the articulatory knowledge and Support Vector Machine (SVM) was used in the classification process. The results of this study are used in clinical applications on dysarthric speech. These two approaches, however, are reliant on neural networks which requires a large amount of data and high processing capability.

Other approaches include modifications to the ASR system components. For example, in [30], modifications to the acoustic model were explored using different convolutional neural networks architectures. An improved HMM was suggested by [31], which is based on automatic error correction for Arabic continuous speech. Another approach was proposed in [32]; it is based on fine-tuning a subset of the neural network layers for non-dysarthric speech models, using speech data obtained from persons with ALS. The neural network architectures that were used are Recurrent Neural Network- Transducer, the Listen, Attend, and Spell. Shahamiri and Salim [33] demonstrated that an improved recognition rate of 24.67% is achieved through the use of a dysarthric multi-network speech recognizer model that employs several artificial neural networks. And, in [34], features were extracted using a convolutive bottleneck network instead of using MFCC.

The limited amount of data on dysarthric speech is one of the greatest challenges in training ASR systems for dysarthric speech. This challenge was approached in [35] using augmented

speech data by employing a virtual microphone array synthesis followed by multi-frame size-based feature extraction. The same challenge of the limited amount of data was approached differently in [36], whereby the ASR system was trained using both dysarthric and non-dysarthric speech data in regard to both the target language and a foreign language. Data from both languages, therefore, was used to build a single acoustic model. The work in [37] suggests that augmentation of dysarthric data can be done by adjusting non-dysarthric speech towards dysarthric speech.

Despite the growing popularity of neural networks, not every problem is solvable using this tool. This thesis explores the differences between dysarthric and non-dysarthric speech and proposes a solution that directly addresses this problem. Contrary to the previous approaches, this work uses two features to build an acoustic model that accommodates dysarthric and non-dysarthric speech, regardless of the inter-speaker variability.

3.3. Articulatory Knowledge

3.3.1.Acoustic-to-Articulatory Inversion

Several researchers have argued that it is possible to employ an acoustic-to-articulatory inversion technique. Sondhi and Gopinath [38] suggest that the cross-sectional area of the vocal tract can be determined by the acoustical measurements at the lips; the solution is based on a set of partial differential equations relating pressure to velocity, while disregarding loss in the vocal tract. Recognizing the exact movement of each of the articulators based on the acoustic signal is considered an ill-posed problem. As there is no unique articulators' configuration for each phoneme; it is possible to produce the same phoneme through different articulators' positions [39]. In [39], the research proposes that this problem can be approached using a mixture density network. Another interesting method to represent the relationship between the articulators and

acoustic signal is suggested by Atal [40], whereby Atal assumes that a non-linear relationship can be represented as linear if a high dimension space is used. Lindblom and Sundberg address the acoustic-to-articulatory inversion more directly [3]. They analyze the effect of articulators' positions on the spectrum of the synthetically generated speech. Specifically, the jaw opening, lips, the body of the tongue, and larynx are studied. The results from [3] are used in this thesis to determine the acoustic features of interest as will be shown in Chapter 4 and Chapter 5.

3.3.2.Recognition Using Articulatory Knowledge

Research shows that techniques based on sound production knowledge have better results in respect to reduced phonemes and word errors [41]. Rudzicz [42] argued that incorporating knowledge of speech production improves the ASR performance quality for people with dysarthric speech. This research resulted in a database of dysarthric articulatory information collected by EMA which is time-aligned with the acoustic data; the database is called TORGO [43]. These results suggest that a dynamic Bayesian network with articulatory knowledge outperform other alternatives. Consequently, an algorithm was developed to estimate articulatory information from the acoustic signal. Another work by Rudzicz [44] explored how the knowledge of speech production, using seven articulatory features, can provide greater improvement in phone recognition. In my thesis work, EMA is not used to obtain articulatory information, but rather an acoustic-to-articulatory inversion process is used instead. In addition, only two articulatory features are used instead of seven features.

Chapter 4

Vowel Recognition Based on Jaw Opening Width

Abstract

The characteristics of the vowels produced by humans are controlled by the positions of speech articulators. This explains the different speech types of a person who speaks non-dysarthric North American English, and a person with a motor function disorder who finds difficulty positioning the articulators and, as such, is classified as a dysarthric speaker. The comprehension of the effect of articulators' positions on the produced vowel will make it possible to build a general classifier of vowels across the two speech types. In this chapter, the relation between the jaw opening width and the first formant (F1) frequency is examined and an F1-based classifier for the vowels of non-dysarthric speech is described. In particular, it is shown that it is possible to recognize five different groups of vowels in non-dysarthric speech using the described F1-based classifier. There is also a discussion as to how to expand the classifier to generalize over dysarthric and non-dysarthric speech.

4.1. Introduction

A sound is a result of an air flow generated in the lungs and passed through a specific vocal tract configuration. Different vocal tract configurations result in different sounds. In cases where vocal cords vibrate while producing a sound, this is called a voiced sound. If the sound propagates for most of the time in the vocal tract and radiates at the lips, this is called non-nasal. A voiced and non-nasal sound which passes through a non-closed vocal tract configuration, is called a vowel.

The configuration set used to produce a certain vowel is controlled mainly by two articulators, namely the jaw and the tongue. Ten vowels will be studied in this chapter; this set is the mutually identified phonemes as vowels in three different resources [4], [8] and [11] as discussed in Chapter 2. The set of ten vowels are: /IY/ as in “Eat”, /IH/ as in “It”, /EH/ as in “Ed”, /AE/ as in “At”, /ER/ as in “Bird”, /AH/ as in “Hut”, /UW/ as in “Two”, /UH/ as in “Hood”, /AO/ as in “Ought”, and /AA/ as in “Odd”.

Since the main difference between dysarthric and non-dysarthric speech is the degree of the control over the articulators, this chapter approaches the problem of recognizing vowels in North American English by identifying the articulators state while producing a certain vowel. Specifically, it focuses on the use of the jaw opening width in recognizing the vowels in both types of speech. A direct way to measure the position of the articulators is through using EMA method as used in the MOCHA-TIMIT database [45] and the TORGO database [43]. However, this is an invasive method. Another approach to measure the articulators’ states is to perform an acoustic-to-articulatory inversion process whereby features from the acoustic signal are translated into articulatory movements. Features extracted from an acoustic signal hold information about the vocal tract’s configuration state. Jaw opening is quantified by the frequency of the first formant (F1) as will be shown in this chapter. As mentioned in Section 2.2.2.2, Formants are defined as the frequency energy groups. Different vocal tract configurations result in different formant frequencies. On a frequency domain representation of a signal, formants are the peaks such as the peak is the center of the formant. On a spectrum, formants are seen as defined intense energy regions.

First, the relation between the frequency of F1 and the jaw opening width is discussed. Then, linear prediction coefficients are used to estimate F1 of a set of samples that represent the

non-dysarthric speech. Finally, an analysis is undertaken as to how to use the jaw opening width to recognize vowels in non-dysarthric speech and whether it can similarly be used to perform the recognition process for dysarthric speech.

4.2. Jaw-First Formant Relation

Lindblom and Sundberg provide an analysis of the effect of articulators' positions on the spectrum of the synthetically generated speech [3]. They showed that F1 is substantially influenced by the mandible movement when all other muscles are stable. They also showed that the effect of tongue height on the frequency spectrum is a combination of the effect of jaw opening and the tongue shape on the spectrum. This means that the tongue height is a secondary feature that can be derived from the vocal tract configuration. For the previously mentioned results from [3], any effect on F1 is considered to be caused by the jaw opening not tongue height as conventionally used. In particular, it was found that the frequency of the F1 centre is correlated with jaw opening. Based on these findings, it can be stated that the jaw-F1 relation is as follows: *the higher the frequency of F1 centre, the wider the jaw opening, and the lower the frequency of F1 centre, the tighter the jaw opening.*

4.3. Methodology

4.3.1. Data Specifications

The dysarthric speech spectrum is divided into two groups. The recorded audio samples were collected from speakers who belong to these two groups. The first group consists of speakers whose first language is English and have non-dysarthric speech. Seven speakers in this group are females and six are males. This group will be referred to as the control group. The second group consists of speakers who have dysarthric speech. One speaker is a female whose English is her first language, one speaker is a male whose English is his first language, and one speaker is a male whose English is his second language. Severity of dysarthric speech varies for the three participants. This group will be referred to as the dysarthric group. The age of the participants in both groups ranges between 20 and 70 years. The data from the control group is used to derive the F1-based classifier. The recorded audio samples are isolated words that contain vowels in North American English. There is a total of ten vowels as discussed in Chapter 2. Each volunteer had a single recording session and was asked to utter ten isolated words: “Eat” containing the vowel /IY/, “It” containing the vowel /IH/, “Ed” containing the vowel /EH/, “At” containing the vowel /AE/, “Hurt”/”Bird” containing the vowel /ER/, “Hut”/“Up” containing the vowel /AH/, “Two” containing the vowel /UW/, “Hood” containing the vowel /UH/, “Ought”/”Dog” containing the vowel /AO/, and “Odd”/”Stop” containing the vowel /AA/. The vowels are then manually extracted from these words for further processing.

The speech signals were obtained using a cardioid condenser microphone. Recording the signal was done using Audacity® recording and editing software v2.2.2.0 [46]. All recordings are in waveform format with the extension .WAV, they are monophonic, sampled at 16 kHz, and

encoded as 16 bits/sample. The speech signals were analyzed using MATLAB_R2019a [47] and plots are generated using both MATLAB_R2019a and Praat [48].

4.3.2. Signal pre-processing

A function called "`audioread ()`" is used to read the .wav recordings and retains the sampled data and the sampling frequency. The part of the sampled data that represents the vowel in the signal is manually extracted. The extracted segment is chosen to represent the longest stable period of the vowel. The reason for not choosing the whole vowel duration is to eliminate the vocal tract changes while producing the previous or next phoneme. This will reduce the coarticulation effect on the recognition of the vowel. This also removes the need for multiple recordings of the same word uttered by the same speaker for the purpose of identifying the characteristics that define a vowel. The extracted segment is then multiplied by a Hamming window of a size similar to the segment length. Signal pre-processing steps are illustrated in Figure 4.1 below.

4.3.3. Estimation of Formants

A function called "`lpc ()`" is used to obtain the linear prediction coefficients that best represent the extracted segment of speech. The example in [49] is followed to estimate the formants. The number of formants that we need to extract is two. The general rule to decide the order of the predictor is to be equal to twice the expected number of formants plus 2 [49]. In theory a linear predictor of the order 8 is sufficient to estimate the 3 formants that are usually observed in a vowel. However, the simulations indicate that a predictor of the order 12 gives the best performance to correctly estimate the formants. This is an empirical result. Thus, the order of the predictor is chosen to be 12. This means that the function will return 12 coefficients of a

polynomial that has a number of roots. Since the coefficients are real-valued, the roots are of the form of complex conjugate pairs. For this, only the roots that have a positive imaginary part sign are kept. The angles corresponding to the retained roots are converted to Hertz. These converted angles represent the frequencies expected to be the frequencies of formants' centres. To decide whether a frequency is a formant frequency or not, the frequency has to pass two conditions. The first is to be higher than 90 Hz, and the second is to have a bandwidth less than 400 Hz, as suggested by [49]. The bandwidth of a frequency is the distance of the root from the unit circle. Those frequencies that pass these two conditions are the estimated formants. The first estimated formant is used to represent the width of jaw opening. The steps of estimating the formants as implemented in Matlab according to [49] is illustrated in Figure 4.1.

In Figure 4.1, the left part of the flowchart shows the processes that need to be done as described earlier, while the right side describes the process to make a decision on what frequencies are considered formants.

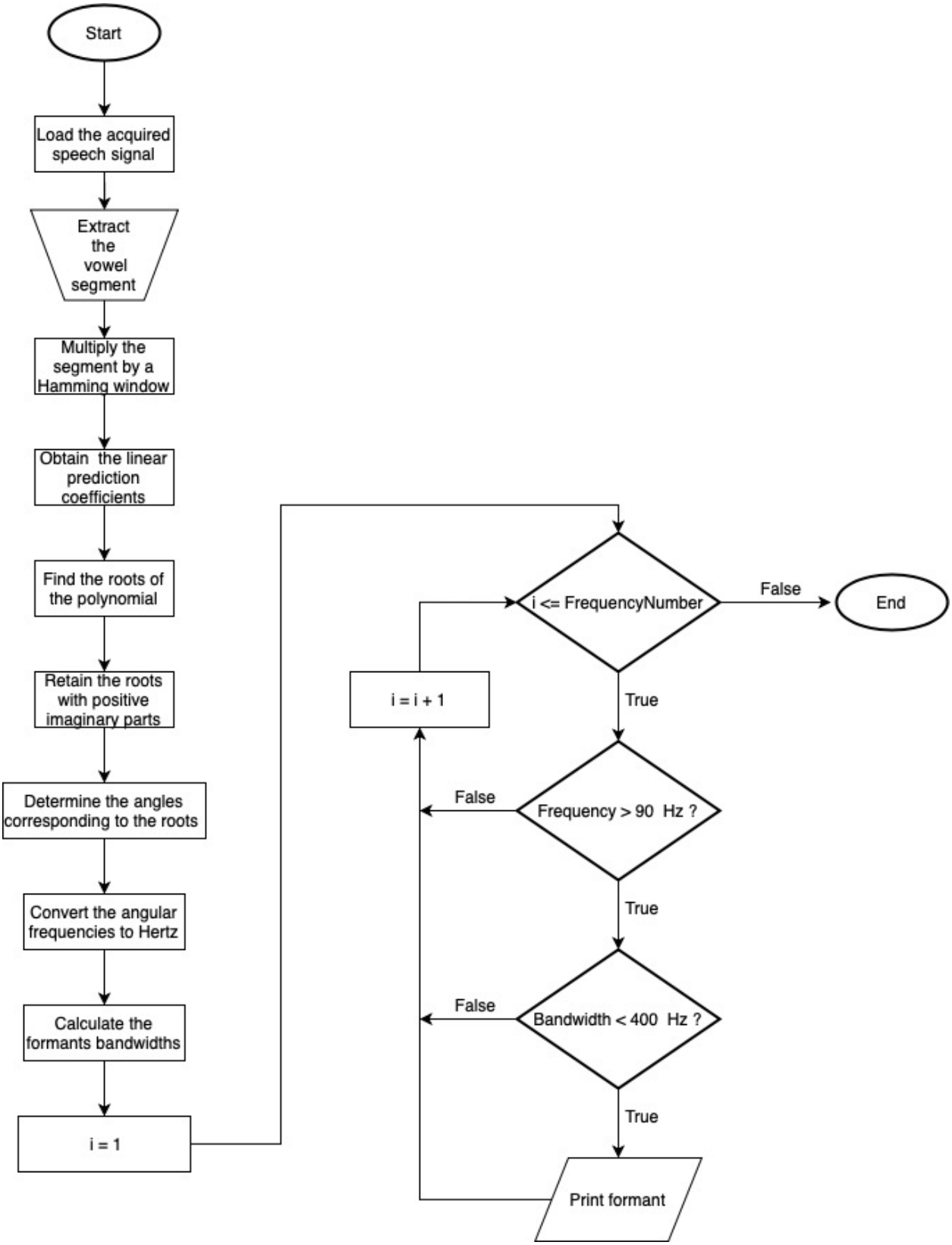


Figure 4.1: Flowchart of the formant estimation process after [49].

4.4. Analysis and Discussion

4.4.1. The Estimated F1 for the Control Group

The centre of F1 is estimated using linear prediction coefficients as mentioned in Section 4.3. Figure 4.2 shows the histograms of the F1 values for the set of ten vowels in North American English extracted from the isolated words uttered by speakers of the control group. Table 4.1 provides the average values for F1 for the set of ten vowels of interest. The mean F1 values in Table 4.1 of a certain vowel are used to indicate the average width of the jaw opening to produce this vowel. For example, the smaller the mean value of F1 for a certain vowel, the tighter the jaw opening. Larger mean values represent wider jaw opening. The order of the average jaw width required for each vowel from tightest to widest is presented in Table 4.1.

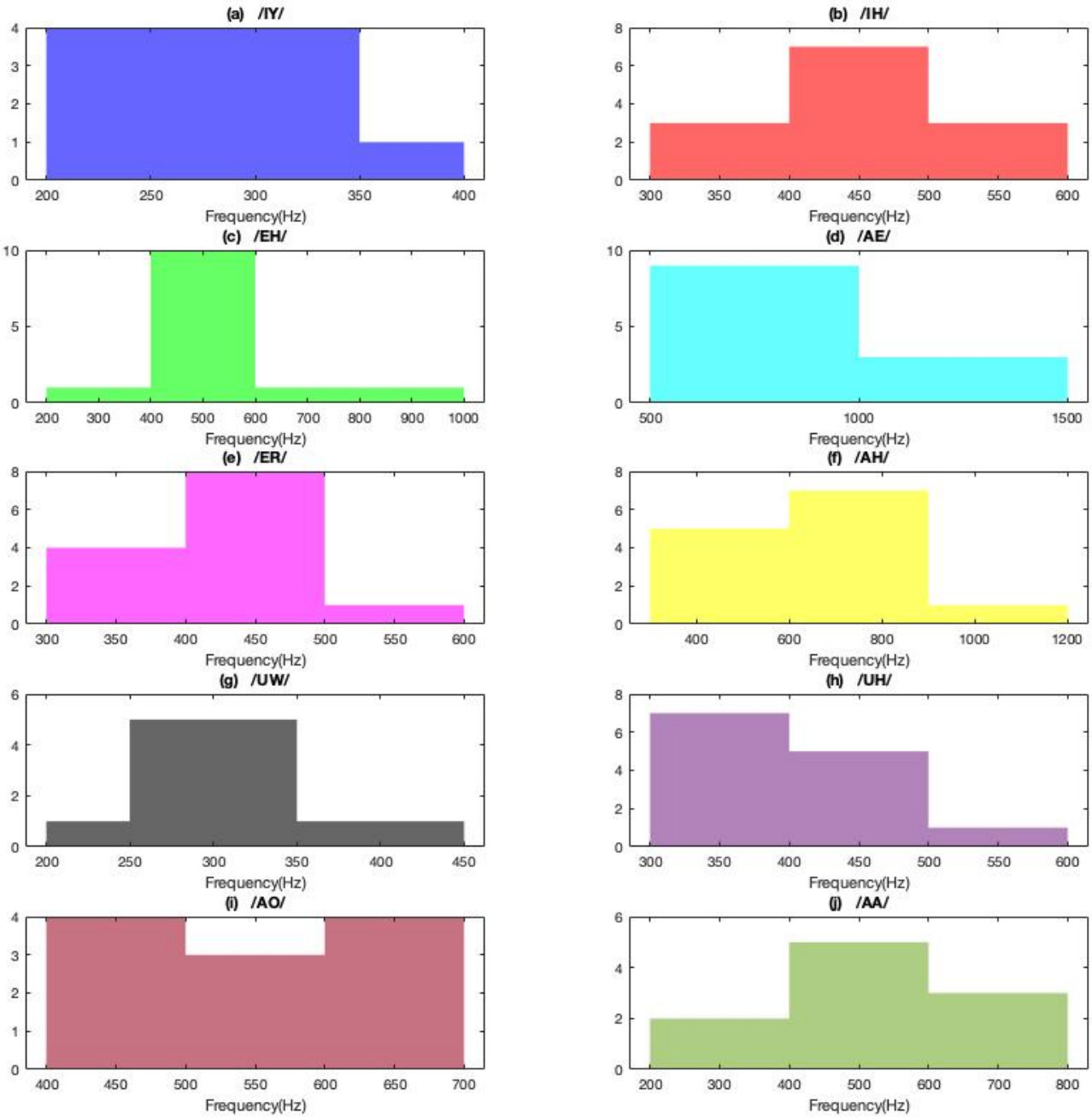


Figure 4.2: Histograms of F1 values for the set of ten vowels in North American English.

Each subplot in Figure 4.2 represents a vowel. The x-axis indicates the F1 frequency, while the y-axis indicates the number of samples that fall in the range of a certain bin.

Vowel	Mean F1 (Hz)	Order of the average jaw width from the tightest to the widest
/IY/	281	1
/IH/	440	5
/EH/	544	8
/AE/	845	10
/ER/	421	4
/AH/	629	9
/UW/	307	2
/UH/	401	3
/AO/	543	7
/AA/	497	6

Table 4.1: The average figures of F1 for the set of ten vowels in North American English for control group.

Each row in Table 4.1 contains a vowel, the vowel's mean F1 frequency in Hz, and a number representing the order of the jaw width, such that 1 represents the tightest jaw opening, and 10 represents the widest jaw opening. The vowel /IY/ which has F1 = 281 Hz ranks first, as it has the tightest jaw opening, followed by the vowel /UW/ which has F1 = 307 Hz, then the vowel /UH/ which has F1 = 401 Hz, the vowel /ER/ which has F1 = 421 Hz, the vowel /IH/ which has F1 = 440 Hz, the vowel /AA/ which has F1 = 497 Hz, the vowel /AO/ which has F1 = 543 Hz, the

vowel /EH/ which has F1 = 544 Hz, the vowel /AH/ which has F1 = 629 Hz, and the vowel /AE/ with the widest jaw opening ranks last and its F1 = 845 Hz.

Because the recorded samples belong to people of wide variety of ages and vocal tract lengths, the produced vowel is different for each individual. This variety in production explains the overlap in histograms of F1values of the vowels as shown in Figure 4.2 and explains the narrow spacing of mean F1 frequencies for some of the vowels as shown in Table 4.1. Thus, it is difficult to recognize a vowel based only on its exact mean F1 frequency. It is possible to reduce the uncertainty in the recognition process if vowels are divided into groups. There are different methods to divide the set of vowels. Nevertheless, it is desired to have a minimum number of groups, which will lead to reducing the computation requirements, allowing to implement the system on machines with the smallest possible processing capabilities. In addition, the number of groups has to take into consideration the unique identification of each vowel in the set. This will become clearer once the features are combined as will be discussed in Chapter 6.

The method chosen to divide the vowels into groups is based on the distance among the F1 mean values. To better depict the distance among mean F1 frequencies, Figure 4.3 visualizes the spacing among the mean F1 values of the considered vowels.

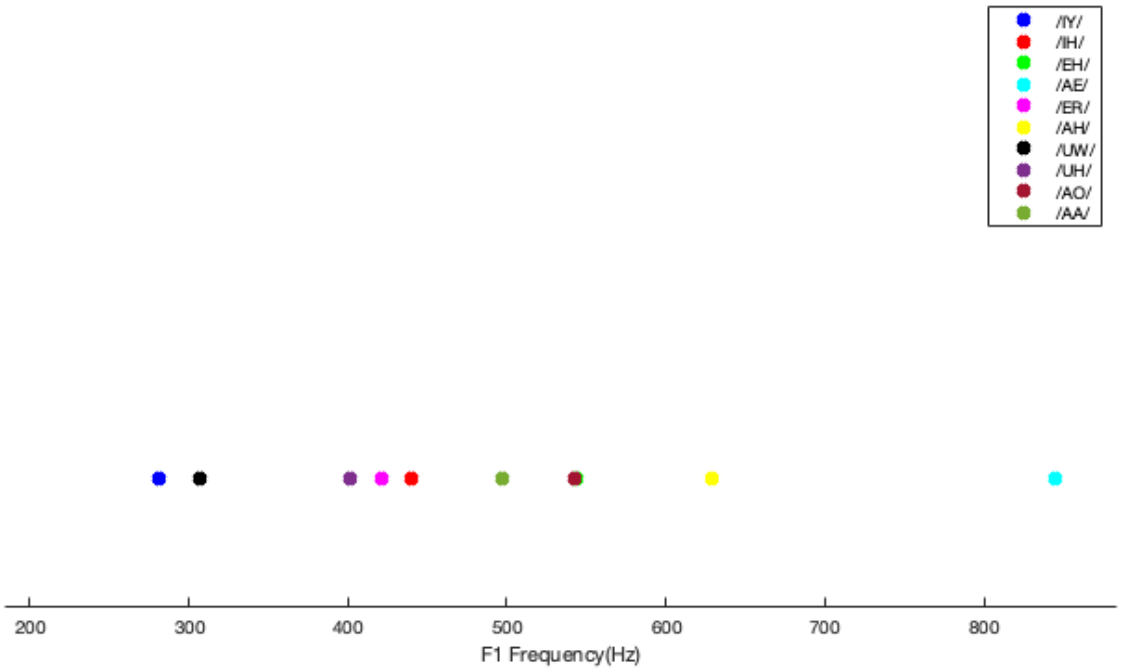


Figure 4.3: Spacing among the average F1 frequency for the set of ten vowels of interest.

In Figure 4.3, the average F1 frequency for vowel /IY/ is the smallest, followed by /UW/, /UH/, /ER/, /IH/, /AA/, /AO/, /EH/, /AH/, and /AE/ which has the largest average F1 frequency.

The first step to group the vowels' mean F1 frequencies into groups is to calculate the distance between adjacent vowels (sorted ascendingly). The distances are as follows: d(IY, UW) = 26 Hz, d(UW, UH) = 94 Hz, d(UH, ER) = 20 Hz, d(ER, IH) = 19 Hz, d(IH, AA) = 57 Hz, d(AA, AO) = 46 Hz, d(AO, EH) = 1 Hz, d(EH, AH) = 85 Hz, d(AH, AE) = 216 Hz.

The second step is to group vowels in between two large distances. In order to choose the least number of groups, the largest four distances are considered. These distances are d(UW, UH), d(IH, AA), d(EH, AH) and d(AH, AE). The four distances divide the whole frequency range into five groups of vowels as follows: the first group contains two vowels /IY/ and /UW/. The second group contains three vowels /UH/, /ER/, and /IH/. The third group contains three vowels /AA/,

/AO/, and /EH/. The fourth group contains a single vowel /AH/. The fifth group contains a single vowel /AE/. Figure 4.4 below depicts the five groups of vowels.

The third step is to designate the range of each of the five groups. Since the lowest accepted frequency for a formant is 90 Hz as mentioned in Section 4.3.3, the lower bound of the first group's range is 90 Hz. The lower bound of all other groups is set to be the F1 frequency of the smallest vowel in that group minus half of the distance from the previous group. The upper bound is set to be the F1 frequency of the largest vowel plus half of the distance from the next group. For example, the upper bound of the first group is the F1 frequency of the vowel /UW/ + 47, which is half of the distance between group I and group II. The range of each group is as follows: group I falls in the range [90, 354] Hz. Group II occupies the range [354, 468.5] Hz. Group III occupies the range [468.5, 586.5] Hz. Group IV falls in the range [586.5, 737] Hz. Group V falls in the range [737, +∞) Hz.

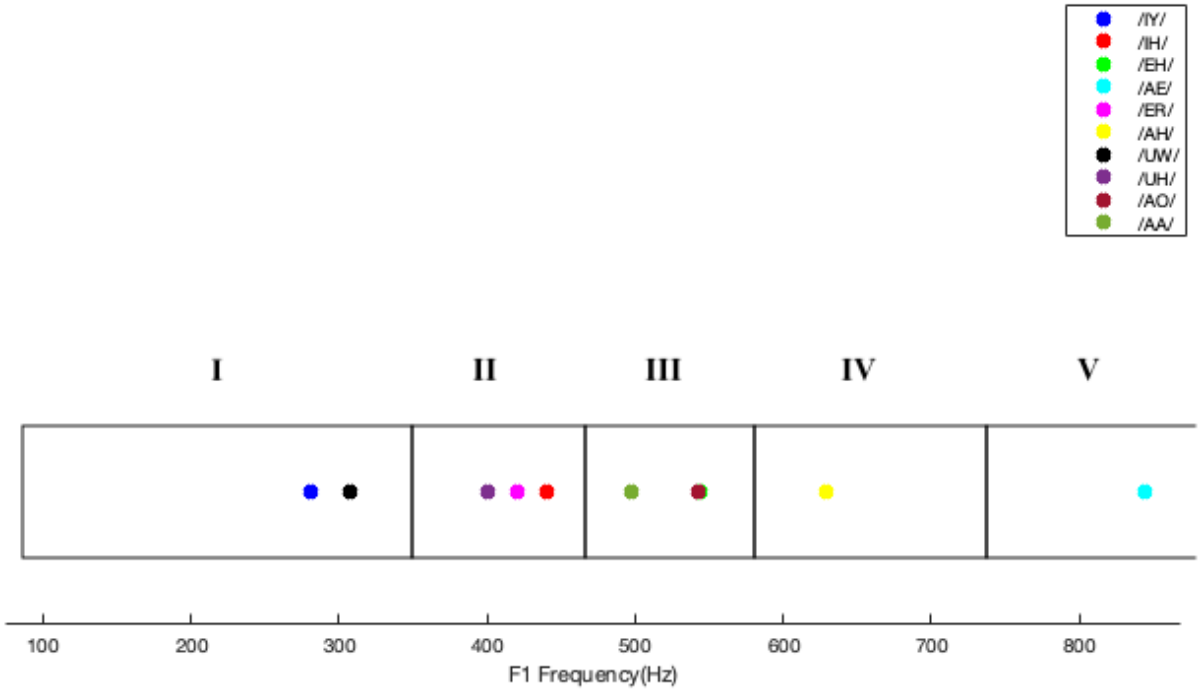


Figure 4.4: Vowels groups according to mean F1 frequency.

Group I contains the vowels /IY/ and /UW/. Group II includes the vowels /UH/, /ER/, /IH/. Group III contains the vowels /AA/, /AO/, /EH/. Group IV contains the vowel /AH/. Group V contains the vowel /AE/.

In the recognition process, the estimated F1 of a vowel recording is checked against each of the five groups of vowels. The group that the estimated F1 frequency falls in represents the group of this vowel. For example, to recognize a vowel that has an estimated F1 frequency of 232 Hz, this number falls in the range of the first group. Thus, the vowel is recognized as either the vowel /IY/ or the vowel /UW/. Groups II and III are more populated, the recognized vowel is one of three vowels. Groups IV and V contain a single vowel each, which means the recognized vowel is final. This suggests that *the knowledge of F1 is good to discern five groups of vowels in non-*

dysarthric North American English. Table 4.2 below provides the F1 group for each vowel in the set of ten vowels of interest.

Vowel	F1 group
/IY/	1
/IH/	2
/EH/	3
/AE/	5
/ER/	2
/AH/	4
/UW/	1
/UH/	2
/AO/	3
/AA/	3

Table 4.2: F1 groups of the set of ten vowels in non-dysarthric speech. Each row contains a vowel and the group according to F1.

The first row contains the vowel /IY/ which belongs to group I of F1. The vowel /IH/ belongs to group II. The vowel /EH/ belongs to group III. The vowel /AE/ belongs to group IV. The vowel /ER/ belongs to group II. The vowel /AH/ belongs to group IV. The vowel /UW/ belongs to group I. The vowel /UH/ belongs to group II. The vowel /AO/ belongs to group III. The vowel /AA/ belongs to group III.

Dividing the set of vowels into five groups according to their mean F1 frequency simplifies the process of indicating the jaw opening width in the production of each vowel. The F1 numerical figures can be seen as a quantified jaw opening in Hertz, and the groups' ranges can be seen as

defining the ranges of various jaw openings in Hertz. The range of possible jaw openings can similarly be divided into five ranges as shown in Figure 4.5 below.

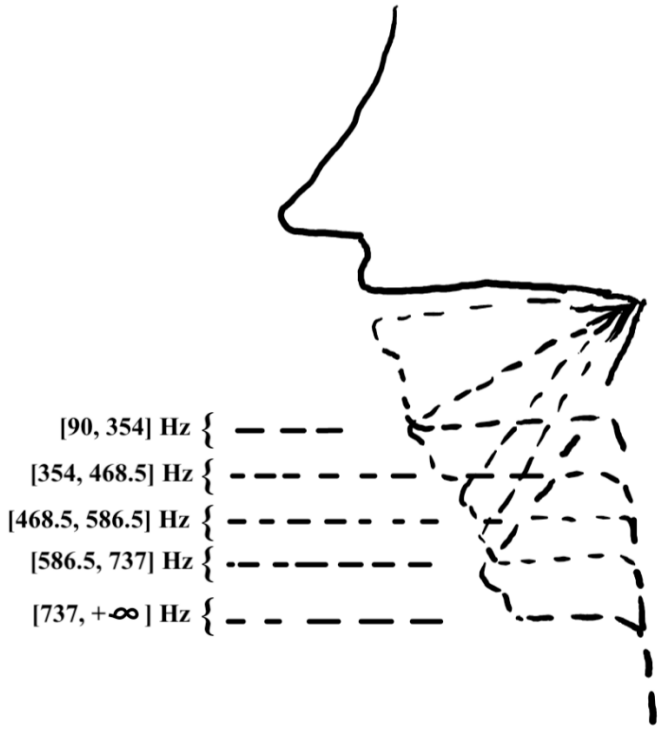


Figure 4.5: Jaw opening ranges to produce the ten vowels of interest and its corresponding F1 frequency ranges.

The five ranges of jaw opening are shown in Figure 4.5. The first range contains the average position to produce the vowels /IY/ and /UW/ and its corresponding F1 frequency range is [90, 354] Hz; the jaw width may take any value in this range. The second range contains the average jaw opening to produce the vowels /UH/, /ER/, and /IH/ and its corresponding F1 frequency range is [354, 468.5] Hz; the jaw width may take any value in this range. The third range contains the mean jaw opening to produce the vowels /AA/, /AO/, and /EH/ and its corresponding F1 frequency range is [468.5, 586.5] Hz; the jaw width may take any value in this range. The fourth range contains the mean jaw opening to produce the vowel /AH/ and its corresponding F1 frequency range is [586.5, 737] Hz; the jaw width may take any value in this range. The fifth range contains

the mean jaw opening to produce the vowel /AE/ and its corresponding F1 frequency range is [737, +∞); the jaw width may take any value in this range.

The recognition of groups of vowels according to jaw opening width works for non-dysarthric speech. However, for individuals with dysarthric speech, jaw opening while producing a vowel may not fall exactly in the same range as the jaw opening of a non-dysarthric speaker. The recognition of dysarthric speech vowels is discussed in the next section.

4.4.2. The Estimated F1 for Dysarthric Group

The classifier described in Section 4.4.1 to recognize the jaw opening of five different groups of vowels based on the mean F1 frequency works for the control group. A similar classifier can be built for the dysarthric group using data from speakers with dysarthria (if this approach is considered, further subdivision of the dysarthric group needs to be done according to the type of dysarthria). However, this will result in an ill posed problem where one-to-many relationship occur as two F1 frequencies that belong to two different ranges may refer to the same vowel group. This makes this approach inconvenient to recognize groups of vowels based on the F1 frequency while accommodating both types of speech. Another approach to build a classifier based on the jaw opening of the set of ten vowels for both types of speech is to combine the samples from both speaker groups into a single group. This approach provides the mean value of F1 frequency ignoring the differences between the two groups of speakers which will result in high error rate. Thus, this approach is inconvenient to recognize groups of vowels based on the F1 frequency.

The F1 frequency of vowels uttered by individuals with dysarthria can exhibit variation. This is because motor function disorders affect individuals differently according to the type of motor function dysfunction which means it affects the control over jaw opening differently. Thus,

the mean F1 frequency in dysarthric group cannot be used to indicate the most common jaw opening to produce a certain vowel for all group members.

Given that the mean F1 frequency of dysarthric group cannot be used to represent the sound production mechanism of all speakers, the jaw-opening-based classifier of the set of ten vowels for both types of speech has to be a customized version of the classifier built for the control group in Section 4.4.1. This customization accommodates the differences in any sound production mechanism. More information is needed to build such a customized classifier. This may be information about other articulators' positions such as tongue and lips or any other feature that conveys a muscle movement. This added information will create redundancy in the classifier that allows for accommodating more different types of speech, while taking into consideration the different abilities to control the articulators responsible for sound production process.

4.5. Conclusion

This chapter presented a classifier for the set of ten vowels in non-dysarthric North American English that is based on the jaw opening quantified by the frequency of F1. The classifier discerns five different groups of vowels based on their mean F1 frequency as uttered by speakers with non-dysarthric speech. However, the jaw opening on its own is not sufficient to recognize vowels in dysarthric speech. This classifier needs to be expanded by adding more features in order to include the vowels in dysarthric speech and make it more robust. The added features have to represent articulatory muscles movement in order to convey the difference in articulation between the two speech types of interest. In the following chapter, the tongue position quantified by the frequency of F2 is studied.

Chapter 5

Vowel Recognition Based on Tongue Longitudinal Position

Abstract

The description of articulators' positions during the production of speech sounds highlights the articulation differences in dysarthric and non-dysarthric speech. Two articulators that contribute to vowel production process are the jaw and the tongue. In this chapter, the tongue position is considered. The relation between the tongue longitudinal position and the F2 frequency is studied and an F2-based classifier for the vowels of non-dysarthric speech is described. Specifically, it is found that it is possible to recognize five different groups of vowels in non-dysarthric speech using the described F2-based classifier.

5.1. Introduction

The vocal tract configuration determines the produced vowel. Jaw opening width contributes to the production process of five different groups of vowels as described in Chapter 4. It is noted, that a classifier based only on F1 is able to recognize only five groups of vowels, but not each vowel independently. Therefore, more features are needed to refine the classification. The added features need to represent articulatory movements. This is because the representation of articulatory movement leads to better discernment of the articulation difference between dysarthric and non-dysarthric speech.

This chapter describes the tongue position feature. The tongue position is quantified by the frequency of the second formant (F2) as will be described later. The linear prediction coefficients are used to estimate F2 of a set of samples that represent the non-dysarthric speech. Finally, an analysis is undertaken as to how to use the tongue position to recognize vowels in non-dysarthric speech and whether it can similarly be used to perform the recognition process for dysarthric speech.

5.2. Tongue-Second Formant Relation

Tongue muscle can be linked to a jelly bag. Since it does not have joints or bones, it is difficult to describe its position using (x, y) or (x, y, z) coordinates. Its complex movement requires a multi-dimensional space to represent the tongue shape, surface area, whether, for example, it is the tongue tip or tongue edges that are moving. In this thesis, the primary interest is in the horizontal position of the tongue.

The paper written by Lindblom and Sundberg [3] provides an analysis of the effect of articulators' positions on the spectrum of speech that is generated synthetically. Lindblom and Sundberg have shown that the shape of the vocal tract can be changed by the tongue hump position whether it is constricted at the back or at the front of the mouth. In particular, it was found that F2 is affected by the tongue position. Based on these findings, it can be stated that the frequency of the F2 centre is correlated with the tongue hump position. That is, *the higher the frequency of F2 centre, the more the front the tongue hump position, and the lower the frequency of F2 centre, the more back the tongue hump position.*

5.3. Methodology

The data used in this chapter is the same as those used in Chapter 4. There are 13 participants in the control group which will be used to derive the F2-based classifier, and 3 participants in the dysarthric group. The age of the participants in both groups ranges between 20 and 70 years. The recorded audio samples are ten isolated words. Each containing a vowel of the set of ten vowels in North American English. The vowels are then manually extracted from these words for further processing. The extracted segment is chosen to represent the longest stable period of the vowel. A linear predictor of the order 12 is used to estimate the second formant of the extracted segment.

5.4. Analysis and Discussion

5.4.1. The Estimated F2 for the Control Group

The centre of F2 is estimated using linear prediction coefficients as previously mentioned in Section 5.3. Figure 5.1 shows the histograms of the F2 values for the set of ten vowels in North American English extracted from the isolated words uttered by speakers of the control group. Table 5.1 provides the average values for F2 for the set of ten vowels of interest. The mean F2 values in Table 5.1 of a certain vowel indicate the average position of the tongue hump to produce this vowel. For example, a small mean value of F2 for a certain vowel means that the tongue hump is positioned in the back of the mouth. Larger mean values represent the case where a tongue hump is positioned in the front of the mouth. The order of the average tongue position required for each vowel from the most back to the most front is presented in Table 5.1.

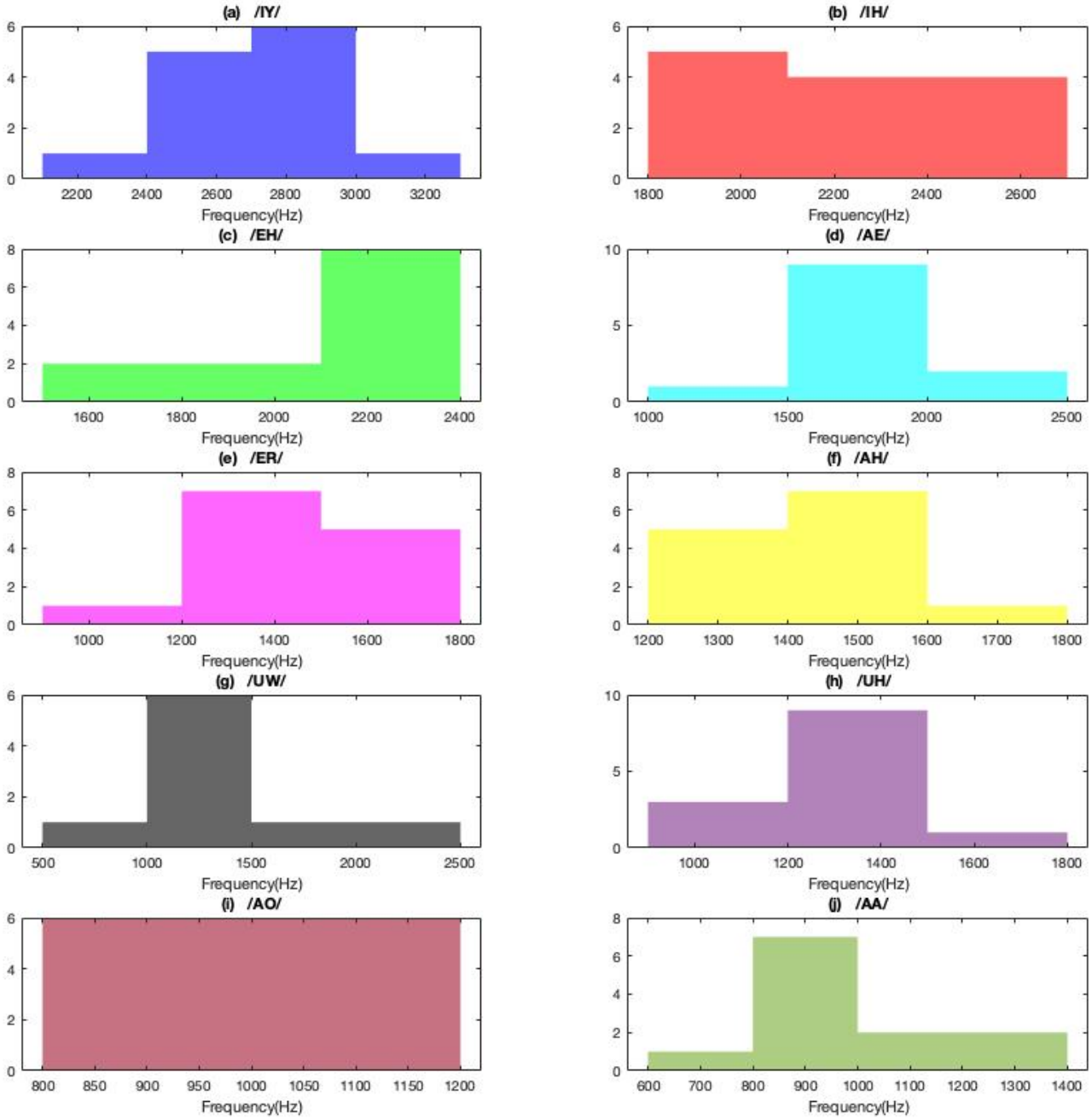


Figure 5.1: Histograms of F2 values for the set of ten vowels in North American English.

Each subplot in Figure 5.1 represents a vowel. The x-axis indicates the F2 frequency, while the y-axis indicates the number of samples that falls in the range of a certain bin.

Vowel	Mean F2 (Hz)	Order of the average tongue position from the most back to the most front position in the mouth
/IY/	2664	10
/IH/	2246	9
/EH/	2084	8
/AE/	1723	7
/ER/	1437	6
/AH/	1419	5
/UW/	1338	4
/UH/	1335	3
/AO/	990	2
/AA/	987	1

Table 5.1: The average figures of F2 for the set of ten vowels in North American English for control group.

Each row in Table 5.1 contains a vowel, the vowel's mean F2 frequency in Hz, and a number representing the order of the tongue position, such that 1 represents the tongue positioned in the most back of the mouth, and ten represents the tongue positioned in the front of the mouth.

The vowel /AA/ which has F2 = 987 Hz ranks first, as it has the most back positioned tongue hump, followed by the vowel /AO/ which has F2 = 990 Hz, then the vowel /UH/ which has F2 = 1335 Hz, the vowel /UW/ which has F2 = 1338 Hz, the vowel /AH/ which has F2 = 1419 Hz, the vowel /ER/ which has F2 = 1437 Hz, the vowel /AE/ which has F2 = 1723 Hz, the vowel /EH/ which has F2 = 2084 Hz, the vowel /IH/ which has F2 = 2246 Hz, and the vowel /IY/ with the most front positioned tongue hump ranks last and its F2 = 2664 Hz.

Because the recorded samples belong to people of wide variety of ages and vocal tract lengths, the produced vowel is different for each individual. This variety in production explains the overlap in histograms of F2 values of the vowels as shown in Figure 5.1, and explains the narrow spacing of mean F2 frequencies for some of the vowels as shown in Table 5.1. Thus, it is difficult to recognize a vowel based only on its exact mean F2 frequency. As discussed in Chapter 4, It is possible to reduce the uncertainty in the recognition process if vowels are grouped in a way that requires low processing power and does not change the unique identification of each vowel as will be discussed in Chapter 6. To better depict the distance among mean F2 frequencies, Figure 5.2 visualizes the spacing among the mean F2 values of the considered vowels.

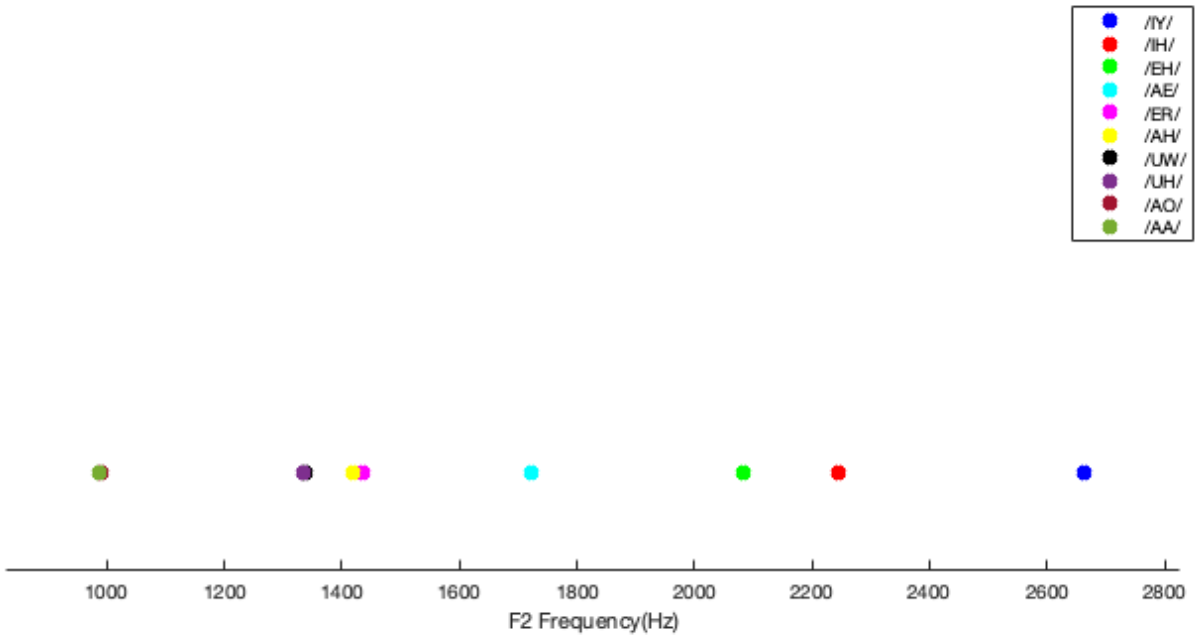


Figure 5.2: Spacing among the average F2 frequency for the set of ten vowels of interest.

In Figure 5.2, the average F2 frequency for vowel /AA/ is the smallest, followed by /AO/, /UH/, /UW/, /AH/, /ER/, /AE/, /EH/, /IH/, and /IY/ which has the largest average F2 frequency.

The first step to group the vowels' mean F2 frequencies is to calculate the distance between adjacent vowels (sorted ascendingly). The distances are as follows: d(AA, AO) = 3 Hz, d(AO, UH) = 345 Hz, d(UH, UW) = 3 Hz, d(UW, AH) = 81 Hz, d(AH, ER) = 18 Hz, d(ER, AE) = 286 Hz, d(AE, EH) = 361 Hz, d(EH, IH) = 162 Hz, d(IH, IT) = 418 Hz.

The second step is to group vowels in between two large distances. As mentioned in Chapter 4, small number of groups is desired. There are four large distances that can be used, these are d(AO, UH), d(ER, AE), d(AE, EH) and d(IH, IY). These four distances divide the whole frequency range into five groups of vowels as follows: the first group contains two vowels /AA/ and /AO/. The second group contains four vowels /UH/, /UW/, /AH/, and /ER/. The third group

contains a single vowel /AE/. The fourth group contains two vowels /EH/ and /IH/. The fifth group contains a single vowel /IY/. Figure 5.3 below depicts the five vowels groups.

The third step is to designate the range of each of the five groups. The lower bound of the first group's range is -∞. The lower bound of all other groups is set to be the F2 frequency of the smallest vowel in that group minus half of the distance from the previous group. The upper bound is set to be the F2 frequency of the largest vowel plus half of the distance from the next group. For example, the upper bound of the first group is the F2 frequency of the vowel /AO/ + 172.5, which is half of the distance between group I and group II. The range of each group is as follows: group I falls in the range (-∞, 1162.5] Hz. Group II occupies the range [1162.5, 1580] Hz. Group III occupies the range [1580, 1903.5] Hz. Group IV falls in the range [1903.5, 2455] Hz. Group V falls in the range [2455, +∞) Hz.

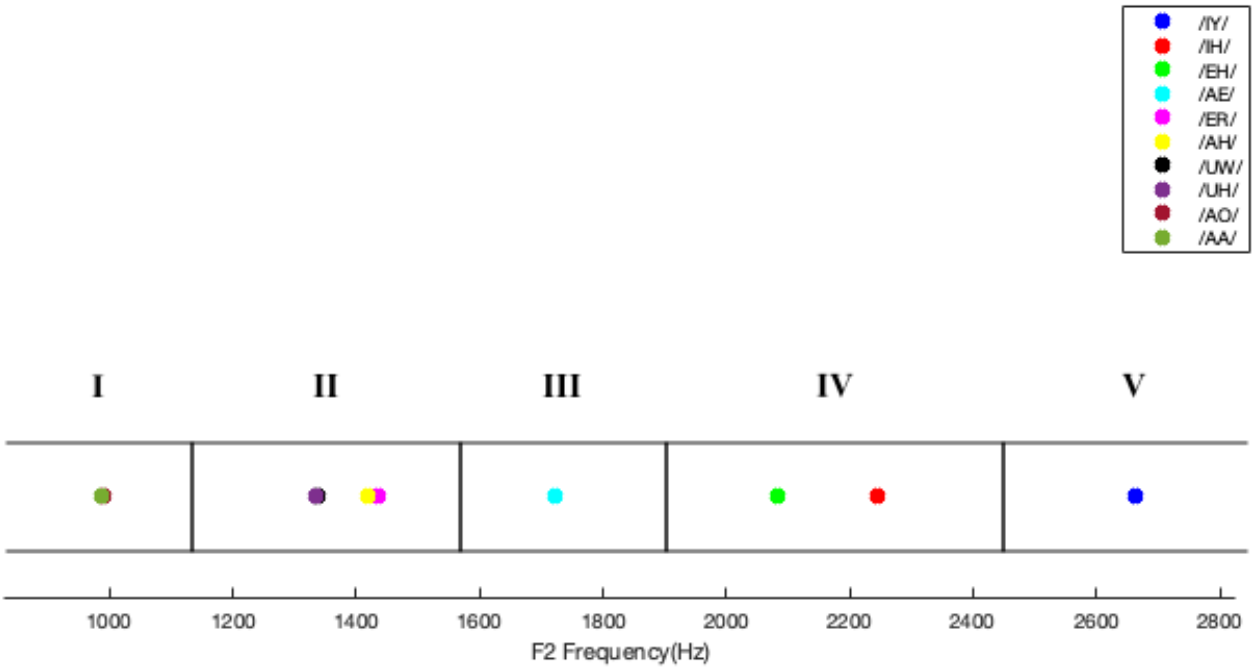


Figure 5.3: Vowels groups according to mean F2 frequency.

Group I on Figure 5.3 contains the vowels /AA/ and /AO/. Group II includes the vowels /UH/, /UW/, /AH/, and /ER/. Group III contains the vowel /AE/. Group IV contains the vowels /EH/ and /IH/. Group V contains the vowel /IY/.

In the recognition process, the estimated F2 of a vowel recording is checked against each of the five groups of vowels. The group that the estimated F2 frequency falls in represents the group of this vowel. For example, to recognize a vowel that has an estimated F2 frequency of 1265 Hz, this number falls in the range of the second group. Thus, the vowel is recognized as either the vowel /UH/, /UW/, /AH/ or the vowel /ER/. Group I contains two vowels, so the recognized vowel is one of two vowels. Group II is more populated, and the recognized vowel is one of four vowels. Group III contains a single vowel, which means the recognized vowel is final. Group IV contains 2 vowels and the recognized vowel is one of 2 options. And group V contains a single vowel, which means the recognized vowel is one option only. This concludes that *the knowledge of F2 is good to discern five groups of vowels in non-dysarthric North American English.* Table 5.2 below provides the F2 group for each vowel in the set of ten vowels of interest.

Vowel	F2 group
/IY/	5
/IH/	4
/EH/	4
/AE/	3
/ER/	2
/AH/	2
/UW/	2
/UH/	2
/AO/	1
/AA/	1

Table 5.2: F2 groups of the set of ten vowels in non-dysarthric speech. Each row contains a vowel and the group according to F2.

The first row in Table 5.2 contains the vowel /IY/ which belongs to group V of F2. The vowel /IH/ belongs to group IV. The vowel /EH/ belongs to group IV. The vowel /AE/ belongs to group III. The vowel /ER/ belongs to group II. The vowel /AH/ belongs to groups II. The vowel /UW/ belongs to group II. The vowel /UH/ belongs to the group II. The vowel /AO/ belongs to group I. The vowel /AA/ belongs to group I.

Dividing the set of vowels into five groups according to their mean F2 frequency simplifies the process of indicating the tongue position in the production of each vowel. The F2 numerical figures are seen as a quantified tongue position in Hertz, and the groups' ranges are seen as defining the ranges of various tongue positions in Hertz. The range of possible tongue positions can similarly be divided into five ranges as shown in Figure 5.4 below.

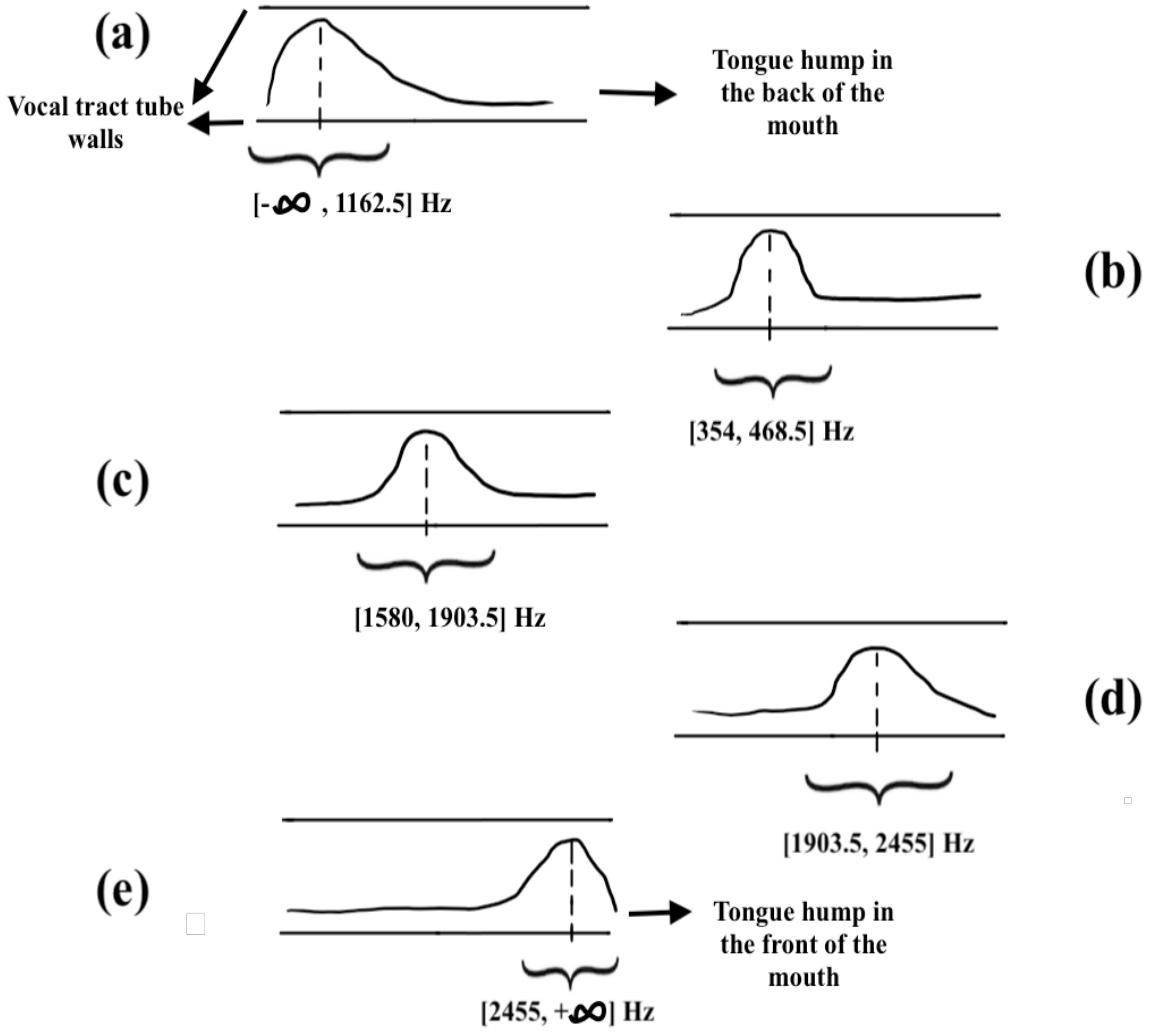


Figure 5.4: Tongue position ranges and its corresponding F2 frequency ranges.

The five ranges of tongue position are shown in Figure 5.4. Each sub plot illustrates the vocal tract tube walls and the tongue hump. (a) The first range is the average tongue position to produce the vowels /AA/ and /AO/ and its corresponding F2 frequency range is (-∞, 1162.5] Hz; the centre of the tongue hump can take any value in this range (this is the most back position range in the mouth). (b) The second range contains the average tongue position to produce the vowels /UH/, /UW/, /AH/, and /ER/ and its corresponding F2 frequency range is [1162.5, 1580] Hz; the centre of the tongue hump can take any value in this range. (c) The third range contains the mean

tongue position to produce the vowel /AE/ and its corresponding F2 frequency range is [1580, 1903.5] Hz; the centre of the tongue hump can take any value in this range. (d) The fourth range contains the mean tongue position to produce the vowels /EH/ and /IH/ and its corresponding F2 frequency range [1903.5, 2455] Hz; the centre of the tongue hump can take any value in this range. (e) The fifth range has the mean tongue position to produce the vowel /IY/ and its corresponding F2 frequency range is [2455, +∞) Hz; the centre of the tongue hump can take any value in this range (this is the most front position range in the mouth).

The recognition of groups of vowels according to tongue position is possible for non-dysarthric speech. However, for individuals with dysarthric speech, the position of the tongue hump while producing a vowel may not fall in the same range as the tongue position of a non-dysarthric speaker. The recognition of dysarthric speech vowels will be discussed in the next section.

5.4.2. The Estimated F2 for Dysarthric Group

Similarly to the results of Chapter 4, the F2-based classifier built in Section 5.4.1 does not accommodate dysarthric speech. Given that the mean F2 frequency of dysarthric group cannot be used to represent the sound production mechanism of all dysarthric speakers, the tongue-position-based classifier of the set of ten vowels for both types of speech has to be a customized version of the classifier built for the control group in Section 5.4.1. More information is needed to build such a customized classifier. This can be information about the jaw opening as discussed in Chapter 4. The combination of classifiers built based on jaw opening and tongue position needs to be studied thoroughly. The use of both features to build a model will contain redundancy that allows for

accommodating more different sounds in speech, while taking into consideration the different abilities to control the articulators responsible for sound production process.

5.5. Conclusion

This chapter presented a classifier for the set of ten vowels in non-dysarthric North American English that is based on the tongue position quantified by the frequency of F2. The classifier discerns five different groups of vowels based on their mean F2 frequency as uttered by speakers with non-dysarthric speech. However, similarly to the results in Chapter 4, the tongue position on its own is not sufficient to recognize vowels in dysarthric speech. This classifier needs to be expanded by adding more features. In the next chapter, the classifier from Chapter 4 and the classifier from this chapter are combined and studied.

Chapter 6

Vowels Quantification into Codewords

Abstract

Acoustic models are built using data from a single type of speech. The comprehension of the effect of articulators' positions on the produced vowel will make it possible to quantify vowels in the spectrum of dysarthric speech into codewords and build an acoustic model that naturally describes different degrees of dysarthria. This chapter presents a set of codewords that identifies all possible vowel and vowel-like sounds that a person can produce with all possible combinations of jaw and tongue positions. The jaw opening width and tongue position are quantified by the first and second formants respectively.

6.1. Introduction

State of the art ASR systems perform poorly when used by a person with dysarthria. The problem has been approached through modifying the architecture of the ANN that was used in building the acoustic model of ASR systems as in [30, 33, 32]. These acoustic models are built to mimic the speech perception mechanisms in humans, but are not customized to accommodate the different speech production mechanisms in dysarthric and non-dysarthric speech. As the unclear articulation is the main identifier of dysarthric speech, the acoustic model needs to include sound production mechanism such that the recognition rate of the ASR system improves. A non-invasive way to collect the articulatory information is by performing an acoustic-to-articulatory inversion process.

In this chapter, it is proposed to replace the acoustic model in conventional ASR systems that is based on ANNs with a simple model that has only two parameters. The proposed model describes all possible vowel and vowel-like sounds a person can produce with all possible combinations of jaw and tongue positions regardless of their speech type. The choice of the two parameters, namely the jaw and the tongue positions, is based on the articulation process as described in chapters 4 and 5.

The rest of the chapter is organized as follows. Section 6.2 provides an overview of the previous work in dysarthric speech recognition area. Section 6.3 provides information about the data used in this study and how features are extracted from the data. In Section 6.4, the vowels characteristics, namely frequency of F1 and F2 are used to build an identification table for each of the ten vowels of interest. The vowel identification table is further analyzed in Section 6.5 to obtain a unique identification table that is used in the recognition process. Finally, Section 6.6 concludes the work in this chapter.

6.2. Related Work

Numerous attempts have been made to improve the performance of ASR systems for dysarthric speech. In general, the problem has been identified as neural-network-related problem. Thus, the solution was either to modify the architecture of the neural network used to build the acoustic model, to add articulatory features to the features fed to the model, or to augment the data of dysarthric speech to obtain enough data to feed into the model.

In this chapter, the problem is identified as an articulation-related problem. Thus, the solution to this problem is not data driven. The suggested solution quantifies all possible sounds that can be produced using all combinations of two articulators instead of focusing on capturing the inter-

speaker variability. Consequently, this solution requires less amount of data and processing power than the power required in neural-network-based solutions that can, for example, be implemented on a small device.

6.3. Data Specifications

The recorded audio samples belong to two different groups of speakers as those used in Chapter 4 and Chapter 5. The first group are speakers whose English is their first language and have non-dysarthric speech. Seven speakers in this group are females and 6 are males. This group will be referred to as the control group and is the group used to derive the model. The second group are speakers who have dysarthric speech. One speaker is a female whose first language is English, one speaker is a male whose first language is English, and one speaker is a male with English as his second language. Severity of dysarthric speech varies for the three participants. This group will be referred to as the dysarthric group and is used to test the model. The recorded audio samples are isolated words that contain vowels in North American English. Each volunteer was asked to utter ten isolated words. The vowels are then manually extracted from these words for further processing. F1 and F2 mean values as estimated in Chapter4 and Chapter 5 are used to derive the model in this chapter.

6.4. Frequency Characteristics of Vowels

Different vowel-specific vocal tract shapes affect the acoustic signal differently. These effects are mainly caused by the jaw opening and tongue position that defines the shape of the vocal tract while producing a certain vowel. The following section is a reminder of the relation

between the frequency characteristics of a vowel and the jaw and tongue positions set to produce this vowel as was mentioned in Chapter 4 and Chapter5.

The paper written by Lindblom and Sundberg [3] provides an analysis on the effect of the articulators' positions on the speech spectrum of synthetic signals. Lindblom and Sundberg have shown that F1 is influenced by the jaw opening when all other muscles are stable. It was also shown that the position of the tongue strongly affects F2.

Based on results from [3], the frequency of the F1 centre is considered to be correlated with jaw opening. Such that the higher the frequency of F1 centre, the wider the jaw opening, and the lower the frequency of F1 centre, the tighter the jaw opening. And the frequency of F2 centre is considered to be correlated with the tongue hump position. Such that the higher the frequency of F2 centre, the more to the front the tongue hump position, and the lower the frequency of F2 centre, the more to the back the tongue hump position.

As mentioned in Chapter 4, there are five different jaw opening ranges that uniquely produce five groups of vowels. And as mentioned in Chapter 5, there are five different tongue positions that are used to uniquely produce five groups of vowels. The ranges of jaw opening width and tongue position are quantified by the F1 and F2 frequency ranges of the vowel groups. Table 6.1 provides the F1 and F2 groups for each vowel as discussed in Chapter 4 and Chapter 5.

Vowel	F1 group	F2 group
/IY/	1	5
/IH/	2	4
/EH/	3	4
/AE/	5	3
/ER/	2	2
/AH/	4	2
/UW/	1	2
/UH/	2	2
/AO/	3	1
/AA/	3	1

Table 6.1: Preliminary F1 and F2 groups of the set of ten vowels in non-dysarthric speech.

Each row in Table 6.1 contains a vowel, the group according to F1, and the group according to F2. The first row contains the vowel /IY/ which belongs to groups 1 and 5 of F1 and F2 respectively. The vowel /IH/ belongs to groups 2 and 4. The vowel /EH/ belongs to groups 3 and 4. The vowel /AE/ belongs to groups 5 and 3. The vowel /ER/ belongs to groups 2 and 2. The vowel /AH/ belongs to groups 4 and 2. The vowel /UW/ belongs to groups 1 and 2. The vowel /UH/ belongs to the groups 2 and 2. The vowel /AO/ belongs to the groups 3 and 1. The vowel /AA/ belongs to the groups 3 and 1.

6.5. Analysis

6.5.1. Unique Vowel Identification

The F1 and F2 groups in Table 6.1 act as vowel identifiers. F1 and F2 groups are combined and used to form a codeword that identifies a vowel. Each codeword consists of two parts, the first is the F1 group and the second is the F2 group. The codewords for the vowels are as follows: /IY/ = 15, /IH/ = 24, /EH/= 34, /AE/ = 53, /ER/ = 22, /AH/ = 42, /UW/ = 12, /UH/ = 22, /AO/ = 31, and /AA/ = 31.

The identification codewords are not unique to each vowel. The vowels /ER/ and /UH/ have the same codeword of 22. According to Yoshida, a teacher in programs to teach English to students whose English is not a first language at the University of California [50], the vowel /ER/ and the phoneme /R/ only differ in timing [51]. As timing is not a trait in dysarthric speech, the phonemes /ER/ and /R/ are considered one gliding semivowel phoneme. Thus, the vowel /ER/ is deleted from the set of vowels of interest. Similarly, the vowels /AO/ and /AA/ have the same codeword of 31. The reason of having the same identification code for these vowels is related to the fact that most North Americans do not distinguish these two vowels [52]. Additionally, in an experiment held to identify vowels by auditory judgment, the pair /AO/ and /AA/ was poorly distinguished [53]. Thus, the vowel /AO/ is deleted from the set of vowels of interest, and the vowel /AA/ is kept. The deletion of vowels /ER/ and /AO/ results in a uniquely identified set of eight vowels. Table 6.2 presents a new set of vowels designed for the purpose of vowels identification taking into consideration dysarthric speech characteristics and the most recent phase of language development.

Vowel	F1 group	F2 group
/IY/	1	5
/IH/	2	4
/EH/	3	4
/AE/	5	3
/AH/	4	2
/UW/	1	2
/UH/	2	2
/AA/	3	1

Table 6.2: Preliminary F1 and F2 groups of the eight vowels in non-dysarthric speech.

Each row in Table 6.2 contains a vowel, the group according to F1, and the group according to F2. The first row contains the vowel /IY/ which belongs to groups 1 and 5 of F1 and F2 respectively. The vowel /IH/ belongs to groups 2 and 4. The vowel /EH/ belongs to groups 3 and 4. The vowel /AE/ belongs to groups 5 and 3. The vowel /AH/ belongs to groups 4 and 2. The vowel /UW/ belongs to groups 1 and 2. The vowel /UH/ belongs to the groups 2 and 2. The vowel /AA/ belongs to the groups 3 and 1.

6.5.2. Representation of the Domain on a 2-D Plane

The total number of all possible codewords in the domain is $5 \times 5 = 25$ codewords because each feature is divided into 5 groups. 8 of these codewords correspond to vowels, and 17 codewords are redundant codewords that correspond to vowel-like sounds. Figure 6.1 depicts the

25 possible codewords on a 2-D plane and specifies which of these codewords correspond to vowels, and which codewords are redundant codewords.

The domain of codewords represents all possible combinations of jaw and tongue positions while producing a sound (quantified by F1 and F2 respectively). The resulting sound of any of these combinations is either a vowel or a vowel-like sound. The vowel codewords represent vowels in non-dysarthric speech, while redundant codewords represent vowel-like sounds. The vowel-like sounds are a result of an attempt from a person with dysarthria to produce a certain vowel, but due to the lack of control over the jaw and tongue, the produced sound is vowel-like.

The recognition of vowels in dysarthric speech based on this domain representation on a 2-D plane can be done in various ways. The recognition is not limited to the method to be discussed next.

An example of the recognition process is the following. The recognition of the exact vowel that the person with dysarthria intended to produce is dependent on the distance between the redundant codeword that represent the vowel-like sound and all neighbouring codewords that represent a vowel. The vowel is recognized as the vowel codeword with the least distance from the redundant codeword. If there are more than one vowel codeword with the same distance from the redundant codeword, a decision cannot be made. Out of the 17 redundant codewords, there are 8 codewords that have a similar distance to at least two different vowels which makes these 8 codewords unusable to recognize a vowel. The last 9 redundant codewords are recognized as follows:

The codewords 11 and 13 are recognized as the vowel /UW/. The codewords 34, 35, 45, and 44 are recognized as the vowel /EH/. The codewords 54 and 55 are recognized as the vowel /AE/. The codeword 51 is recognized as the vowel /AH/.

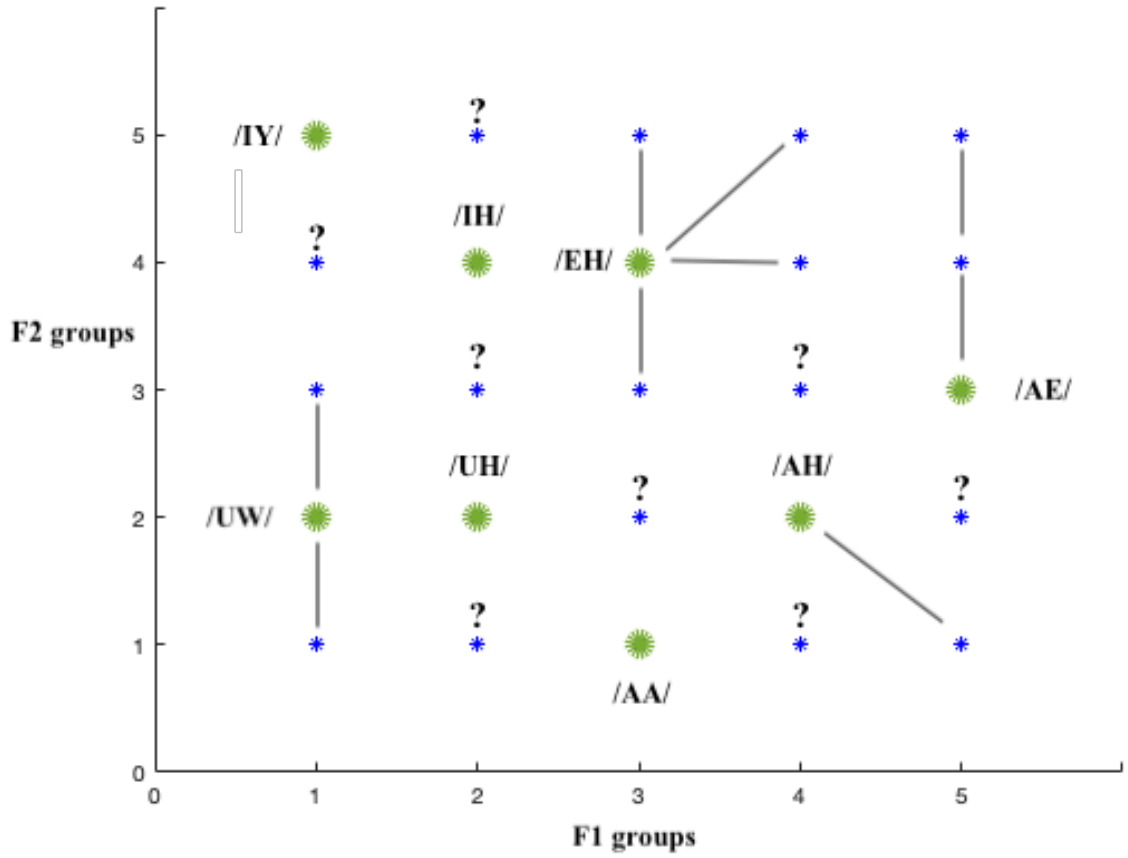


Figure 6.1: All possible codewords of two features divided into 5 groups each.

The x-axis in Figure 6.1 represents the groups of F1. The y-axis represents the groups of F2. The coordinates of each point on the 2-D plane form a codeword. The bold points are the vowels' codewords. The points with a question mark on top of them are redundant codewords where no decision can be made, and the redundant codewords connected to a vowel codeword with a line are the codewords that can be recognized as the connected vowel.

6.5.3. Use Case

To further demonstrate how this domain of codewords can be used in the recognition of dysarthric speech, the data from three different speakers with dysarthria are analyzed. Three speakers: speaker #10, speaker #12, and speaker #13 -with different degrees of dysarthria- were asked to utter the word “Ed”. The vowel /EH/ is extracted from the signal and the frequencies of F1 and F2 are estimated, then a codeword is created. The codeword representing the vowel /EH/ uttered by speaker #10 is 22. The codeword representing the same vowel uttered by speaker #12 is 34. And the codeword representing the vowel /EH/ uttered by speaker #13 is 44. As defined in the domain of codewords in Table 6.2, the vowel /EH/ is represented by the codeword 34. In the case of speaker #10, the codeword 22 will be recognized as the vowel /UH/. The speaker intended to utter the vowel /EH/ but the place of the jaw and tongue were positioned for the generation of a completely different vowel. In the case of speaker #12, the codeword is successfully recognized as the vowel /EH/. In the case of speaker # 13, the codeword 44 corresponds to a redundant codeword that is recognized as the vowel /EH/, so the vowel is successfully recognized.

6.5.4. Domain Analysis

The codewords of the set of 8 vowels are examined for possible refinements. The goal is to have a domain of codewords where most of the codewords are used and the number of redundant codewords is minimal. This includes examining the number of features (Such as F1 and F2) and number of groups in each feature to create the desired domain of codewords.

Theoretically, 8 different vowels are uniquely identified using 1 feature if this feature is divided into 8 groups, such that each group contains a single vowel only. There are a total number of 8 codewords in the domain. This is an optimal case where no codeword is redundant. The same

8 vowels can be uniquely identified using 2 features if each feature is divided into 3 groups. This is because $\lceil log_3 8 \rceil = 2$. 2 is the number of features, the log base is the number of groups (size of alphabet), and 8 is the number of vowels. Using 2 features divided into 3 groups will result in 9 unique codewords. As we only have eight vowels, one codeword is redundant and will not be used. An optimal case will result from dividing one of the features into 2 groups, and the other into 4 groups. The total number of codewords in the domain would be $2 \times 4 = 8$ codewords.

Practically, 8 different vowels are not uniquely identified using a single feature as shown in Chapter 4 and Chapter 5. However, 8 vowels are uniquely identified using 2 features as shown in Table 6.2. Although it was suggested earlier that an optimal case will result from dividing one feature into 2 groups and the other into 4 groups, this is not feasible in the case of identifying vowels because these features describe the natural physical movement of jaw and tongue to produce sounds. Each of the two features is divided into 5 groups to convey the natural process of producing each vowel.

It is preferable to have fewer groups as this decreases the precision degree needed to uniquely identify the vowels. Fewer groups also accommodate more different speech mechanisms by decreasing the number of redundant codewords where no decision can be made. Combining adjacent groups in a feature is examined to explore possible domain of a smaller size than 25 unique codewords.

The first feature (F1): combining groups 1 and 2 of F1 feature into a single group 1 will result in two vowels that have the same codeword. The vowel /UW/ is represented by the codeword 12, and the vowel /UH/ will have the codeword 12 if groups 1 and 2 are combined. Since combining groups 1 and 2 changes the uniqueness of the codewords, groups 1 and 2 cannot be combined.

Combining groups 2 and 3 of F1 feature into a single group 2 will also result in two vowels having the same codeword. The vowel /IH/ has the codeword 24, and the vowel /EH/ will have the codeword 24 if groups 2 and 3 are combined. Thus, groups 2 and 3 of the first feature cannot be combined.

Combining groups 3 and 4 of F1 feature into a single group 3 does not change the uniqueness of any of the codewords. And combining groups 4 and 5 into a single group does not change the uniqueness of the codewords either. For that, groups 3, 4, and 5 of the first feature are combined into a single group 3. The final number of groups that F1 feature is divided into is 3 groups. Table 6.3 below shows the new codewords set after combining groups 3, 4, and 5 in F1 feature.

Vowel	F1 group	F2 group
/IY/	1	5
/IH/	2	4
/EH/	3	4
/AE/	3	3
/AH/	3	2
/UW/	1	2
/UH/	2	2
/AA/	3	1

Table 6.3: The new vowel codewords after F1 groups combination. The codewords of the vowels /AE/ and /AH/ have changed into 33 and 32 respectively.

The second feature (F2): In the new codeword set of Table 6.3, combining groups 1 and 2 of F2 feature into a single group 1 will result in two vowels that have the same codeword. The vowel /AA/ is represented by the codeword 31, and the vowel /AH/ will have the codeword 31 if groups 1 and 2 are combined. Thus, groups 1 and 2 cannot be combined.

Combining groups 2 and 3 of F2 feature into a single group 2 will also result in two vowels having the same codeword. The vowel /AH/ has the code word 32, and the vowel /AE/ will have the codeword 32 if the groups 2 and 3 are combined. Thus, groups 2 and 3 of the second feature cannot be combined.

Combining groups 3 and 4 of F2 feature into a single group 3 will result in two vowels represented by the same codeword. The vowel /AE/ has the code word 33, and the vowel /EH/ will have the codeword 33 if the groups 3 and 4 are combined. Thus, groups 3 and 4 of the second feature cannot be combined.

Combining groups 4 and 5 of F2 feature into a single group 4 does not change the uniqueness of any of the codewords. Thus, groups 4 and 5 of the second feature are combined into a single group 4. The final number of groups that F2 feature is divided into is 4 groups.

Since no further refinements can be done on any of the two features, these results are considered final. Table 6.4 below shows the final codewords set after combining groups 3, 4, and 5 in F1 feature, and groups 4 and 5 in F2 feature.

Vowel	F1 group	F2 group
/IY/	1	4
/IH/	2	4
/EH/	3	4
/AE/	3	3
/AH/	3	2
/UW/	1	2
/UH/	2	2
/AA/	3	1

Table 6.4: The final vowel codewords after groups combination in both features F1 and F2.

The final domain of codewords is of the size $4 \times 3 = 12$. Figure 6.2 depicts all possible codewords on a 2-D plane. This domain represents the vowels and vowels-like sounds in North American English regardless of the speech type if groups inside F1 and F2 features are combined. 8 codewords out of the 12 codewords correspond to vowels, and 4 codewords are redundant codewords. It is clear that the combination of groups inside features increased the margins of vowels and reduced the number of redundant codewords. 1 out of the 4 redundant codewords is recognizable as a vowel. This is the codeword 11 recognized as the vowel /UW/. The rest 3 codewords (13, 21, and 23) have a same distance to more than one vowel, which makes these redundant codewords unrecognizable. The uncertainty in the whole domain has been reduced from 8 unrecognizable codewords to 3 unrecognizable codewords. Nevertheless, if the features chosen to represent the vowels are features that are either hard to be extracted or do not precisely represent

the vowel, error rate will increase with the vowels margins increase. Thus, the choice of merging groups in a certain feature or not is dependent on the features set.

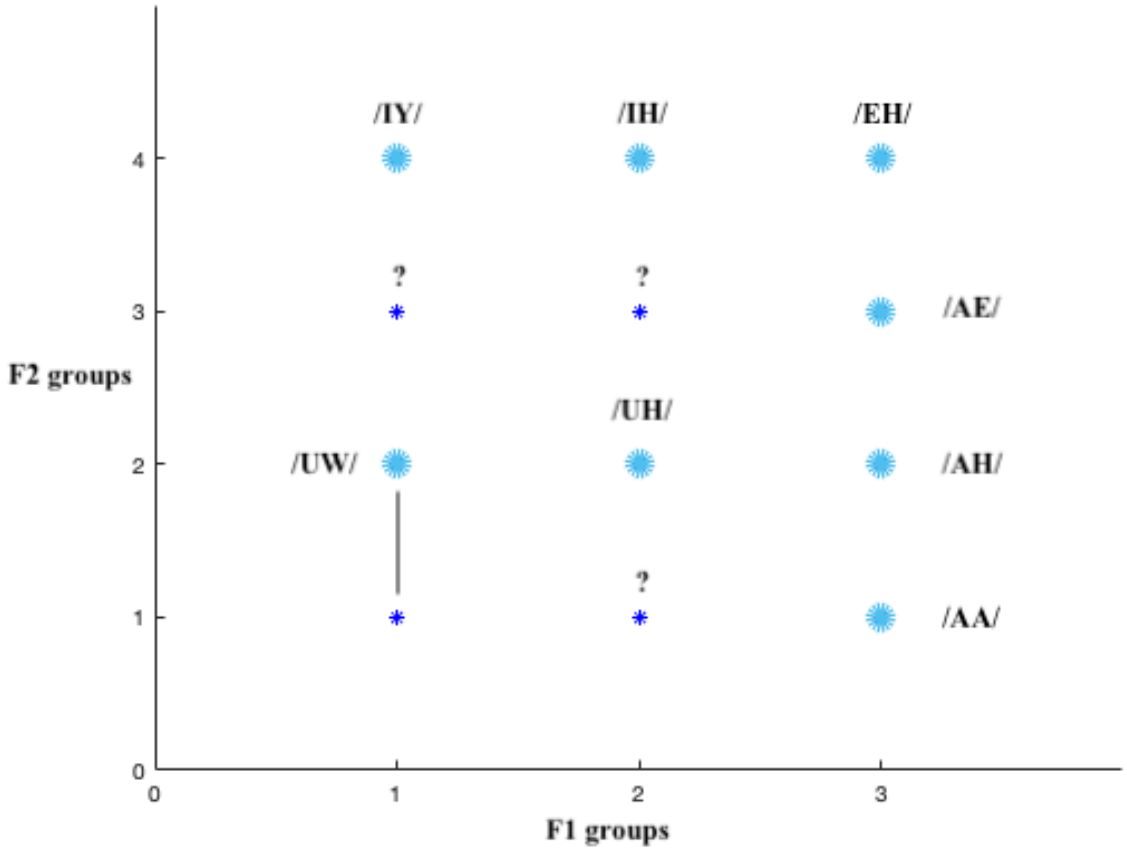


Figure 6.2: All possible codewords of F1 divided into 3 groups and F2 divided into 4 groups.

The x-axis in Figure 6.2 represents the 3 groups of F1. The y-axis represents the 4 groups of F2. The coordinates of each point on the 2-D plane is a codeword. The bold points are vowels codewords. The points with a question mark on top of them is a redundant codeword where no decision can be made, and the redundant codewords connected to a vowel codeword with a line are the codewords that are recognized as the connected vowel.

6.6. Dysarthric Speech Recognition Based on Vowels

Once vowels are recognized based on their estimated F1 and F2 frequencies, the process of recognizing other phonemes in the language can be planned. The first step in the plan is to use the information found in the vowel spectrum as it holds important information about adjacent phonemes. Part of the vowel spectrum duration represents the jaw and tongue position that are set to produce the vowel. The rest of the vowel spectrum represents the changes in the tongue and jaw positions while preparing to produce the next phoneme or while preparing to produce the vowel. An example of the effect of adjacent phonemes on the vowel is shown in Figure 6.3 below.

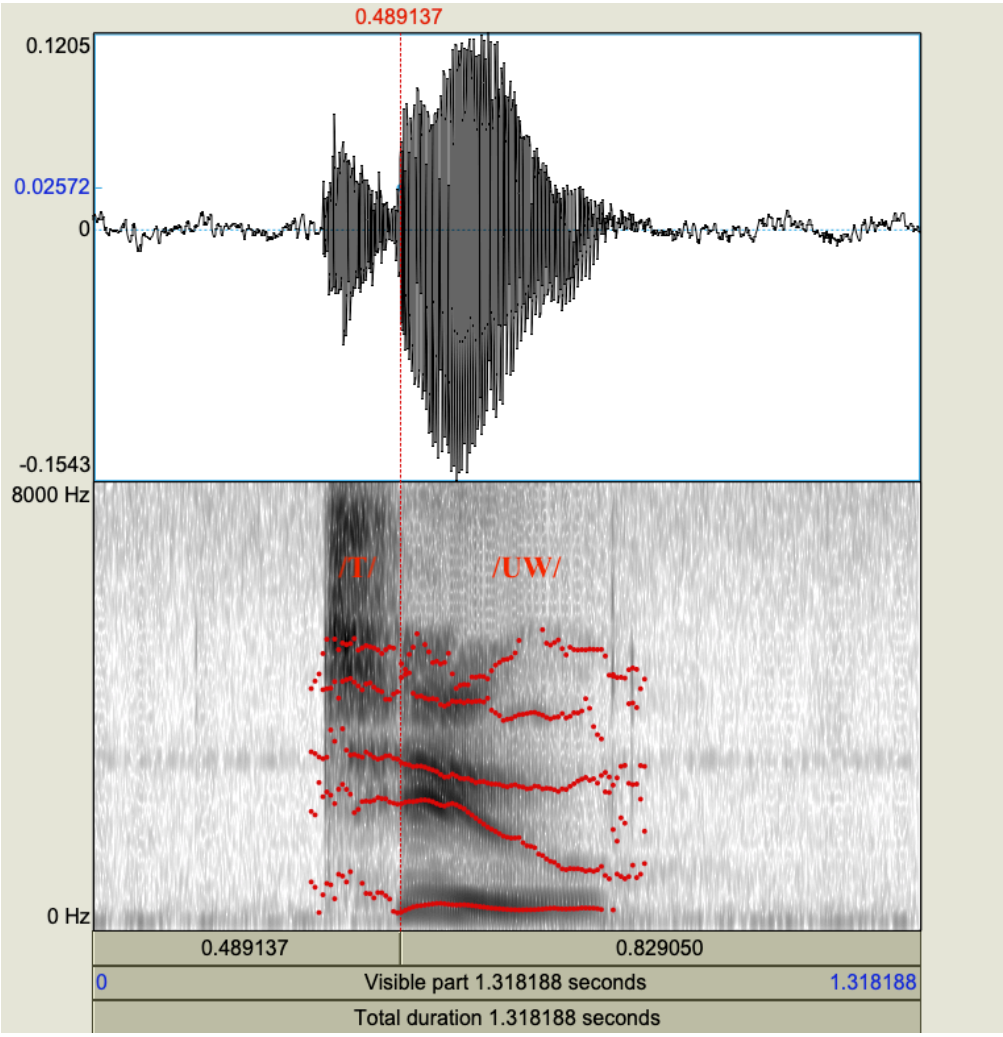


Figure 6.3: The word "Two" uttered by speaker #4.

Speaker #4 is a female speaker who speaks English as a first language and has non-dysarthric speech. The top of Figure 6.3 shows the waveform of the signal and the spectrum is below it. The spectrum is divided into two parts. The first part corresponds to the phoneme /T/. The second part corresponds to the phoneme /UW/. The horizontal red dotted line represents formants. There are five obvious formants in this spectrum of the vowel /UW/.

The first part of the spectrum is a plosive sound where no sound is radiating in the mouth. Thus, the formants are not clear in this part. To produce the plosive /T/ the jaw opening has to be tight and the tongue has to be positioned in the front of the mouth. The jaw also has to be opened tightly to produce the vowel /UW/, but the tongue has to be in the back of the mouth. The spectrum narrates the process of producing the word “Two”. First, the jaw is tight, and the tongue is in the front to produce the /T/ sound. After producing the sound /T/, the closure point is released, and the sound radiates freely in the mouth cavity which explains the presence of formants on the spectrum before the sound /UW/ is heard. The beginning of formants appearing on the spectrum represent the state of a tight jaw and front tongue then immediately changes to a little wider jaw opening (slightly higher F1 value). This change is explained as the tongue is changing its shape (not position), and in order to change the shape, the jaw opens wider to allow extra space. Then, the jaw is set back to the required position to produce the vowel /UW/, and the tongue gradually moves towards the back of the mouth to the required position to produce the vowel /UW/ (F2 centre starts high then it dips down gradually as the tongue is moving back to the required position).

To use information about vowels in the recognition of the whole speech signal, the behaviour of F1 and F2 needs to be studied in whole throughout the vowel period instead of just obtaining the mean value of F1 and F2 of the stable vowel segment period. This will enable us to study the progress of the articulators’ movement which will define the progress of the acoustic

signal. In order to do this, there is a need to build a robust algorithm that is sensitive to the slightest change in F1 and F2 frequencies. Furthermore, knowledge of the sound production mechanism of other phonemes is also needed to categorize the different effects of articulators' movements while producing a phoneme on the vowel's spectrum due to coarticulation. This step requires extensive study in order to identify the best set of features that describes the articulation process of the rest of the phonemes, and to identify how each phoneme affects each of the eight vowels. Once we have more information about the rest of the phonemes, we will have a better idea on how to build a vowel-based ASR system for dysarthric speech based on F1 and F2 frequencies of eight vowels in North American English.

6.7. Conclusion

This chapter presented a model for the set of eight vowels and seventeen vowel-like sounds in North American English that is based on all possible combinations of jaw opening quantified by the frequency of F1 and on the tongue position quantified by the frequency of F2. The sounds are assigned unique codewords based on F1 and F2. The work has shown that a human with a jaw and a tongue can only produce a certain number of sounds. Thus, the problem is quantified. The number of features and the number or groups each feature is divided into is a trade-off between the required degree of precision and the available processing power. Multiple ways, other than the one discussed in this chapter, can be used to perform the recognition process. And multiple models can be built based on the proposed model in this chapter.

Chapter 7

Concluding Remarks

The purpose of this thesis was to explore the existence of *a robust model of vowels and vowel-like sounds in North American English based on all possible natural combinations of jaw and tongue positions, that allows for accommodating dysarthric speech regardless of the inter-speaker variability and requires minimum processing power.* This model was proven to exist through the use of features that convey the natural process of sound production. The first formant was used to quantify the jaw opening width as shown in Chapter 4, and the second formant was used to quantify the tongue position as described in Chapter 5. Each vowel was quantified by a two-digit codeword. The first digit represented the jaw position, while the second digit represented the tongue position. The domain of all possible codewords was used to represent the domain of all possible sounds produced by a person who has a jaw and a tongue regardless of their speech type or personal variability. A summary of the contributions made by this study and future work are described in the next sections.

7.1. Summary of Contributions

Several contributions have been made in this research. The most prominent contributions are as follows:

- The construction of a new vowels set in North American English specifically designed for dysarthric speech taking into consideration the language developments, shown in Section 6.5.1.

- The acoustic model represented by the domain of codewords that quantify all vowels and vowel-like sounds in North American English that can be produced by all possible combinations of jaw and tongue positions, which accommodates dysarthric speech, suggested in Section 6.5.2, and demonstrated through the use of real speech samples acquired from individuals with dysarthria, shown in section 6.5.3.

7.2. Future Work

In this thesis, a new concept was proposed to solve the problem of dysarthric speech recognition. Consequently, more work and exploration would be expected to follow. Future work will depend on the purpose of the application where this solution is used. To use this solution in applications that accommodate accented English, word-level recognition can be used to decrease the error rate. It is common for people who speak English accents other than North American to replace one vowel with another in some words, which is an undetectable error. One way to detect this error is to perform phoneme-level recognition then word-level recognition as a second step. If this solution is to be used in an application to recognize a different language, the set of features should represent the articulators' state of the speakers of that language. The number of features may be different, and the number of groups that each feature is divided into may also be different, but the main concept of numerical representation of vowels is transferable across languages. The usage of this solution in clinical applications, such as monitoring the development of motor function disorders or the severity level of dysarthria is also possible.

Several parts of this research can be developed into independent research areas that contribute to improvements in the dysarthric speech recognition research area. An interesting topic to explore further would be the possibility of deriving an equation that describes the tongue height

in terms of jaw opening. This may lead to un-correlating articulators' movement, a well-known problem, which will enable a more precise indication of the articulators' states and a better acoustic-to-articulatory inversion process. In turn, this would identify all contributors to the first formant height. Another part of this thesis to continue investigating is whether or not a better set of features exists to better represent the different articulation of each vowel. In this thesis, F1 and F2 were used to represent the jaw and tongue positions. There may be other time-domain or frequency-domain features that represent the articulators' states more precisely. Similarly, an investigation should be conducted on the number of groups in each feature needed to create a code with the lowest error rate. The recognition process is also a field that can be further investigated. The principles that apply to communication theory also apply to the set of sounds if we consider the vowels set as the set of transmitted messages and the whole domain of codewords as the set of received messages. The most practical topic in this thesis to be further developed and used in state-of-the-art ASR systems is to build a vowel-based ASR system as discussed in Section 6.6. An extensive study and several steps need to be performed in order to be able to use the co-articulation phenomena (once considered a problem) to recognize the whole speech signal. Further improvements can be explored through the study of combining HMM or ANN with the model proposed in this thesis.

It is true that ANNs has been leading the research of ASR for the past decade or so. Yet, this approach requires costly resources like high processing capabilities and a large amount of data. In contrast, the approach used in this thesis did not require high processing power or large amounts of data. The solution has been made possible using a blend of knowledge from signal processing, information theory, linguistics, and disability rights.

References

[1] F. L. Darley, A. E. Aronson, J. R. Brown, "Differential Diagnostic Patterns of Dysarthria", *J. of Speech lang. Hear. Res.* Vol. 12, no. 2, pp. 246-269, 1969.

[2] F. Rudzicz, "Production Knowledge in the Recognition of Dysarthric Speech", ch. 2, Ph.D. dissertation, Dept. Comput. Sci., Univ. of Toronto, Ontario, Canada, 2011.

[3] B. Lindblom, J. Sundberg, "Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement" *J. of the Acoust. Soc. of America,* vol. 50, pp.1166-1179, 1971.

[4] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, 2nd Ed., Springer-Verlag, New York, 1972.

[5] J. Harington and S. Cassidy, "The Scope of Speech Acoustics", in *Techniques in Speech Acoustics,* Dordrecht, The Netherlands: Kluwer Academic Publishers, 1999, pp. 1-8.

[6] G. M. Murray, "Jaw Movement and Its Control" in *Functional Occlusion in Restorative Dentistry and Prosthodontics*, Mosby, 2016, PP 55-66.

[7] D. O'Shaughnessy, "Speech Production and Acoustic Phonetics" in *Speech Communications – Human and Machine*, 2nd Ed., New York, NY, USA: IEEE Press 2000, pp 47.

[8] L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signals,* Upper Saddle River, NJ, USA: Prentice-Hall, 1978.

[9] Phoneme. (n) In Merriam-Webster's collegiate dictionary. merriam-webster.com. https://www.merriam-webster.com/dictionary/phoneme (Accessed Mar. 10, 2020).

[10] A. L. Bizzocchi, "How Many Phonemes Does the English Language Have?", *Int. J. on Stud. in English Lang. and Literature*, vol. 5, no. 10, pp. 36-46, 2017.

[11] K. Lenzo. "The CMU Pronouncing Dictionary". CMU.edu. http://www.speech.cs.cmu.edu/cgi-bin/cmudict/ (Accessed Mar. 10, 2020).

[12] F. L. Wightman, "Pitch Perception: An Example of Auditory Pattern Recognition" in *Auditory and Visual Pattern Recognition*, D. J. Getty and J. H. Howard Jr., Eds., Hillsdale, NJ, USA: Lawrence Erlbaum Associates,1981, pp. 3-25.

[13] J. T. Tou, "A Feature-Extraction Approach to Auditory Pattern Recognition" in *Auditory and Visual Pattern Recognition*, D. J. Getty and J. H. Howard Jr., Eds., Hillsdale, NJ, USA: Lawrence Erlbaum Associates,1981, pp. 129-142.

[14] J. H. Howard, Jr., J. A. Ballas, "Feature Selection in Auditory Perception" in *Auditory and Visual Pattern Recognition*, D. J. Getty and J. H. Howard Jr., Eds., Hillsdale, NJ, USA: Lawrence Erlbaum Associates,1981, pp. 181-195

[15] D. B. Pisoni and J. R. Sawusch, "Some Stages of Processing in Speech Perception" in *Proc. Symp. dynamic aspects of speech perception,* in Structure and Process in Speech Perception, A. Cohen and S. G. Nooteboom, Eds. Eindhoven, Netherlands: Springer-Verlag Berlin Heidelberg New York, 1975, pp. 16-35.

[16] S. G. Nootboom and A. Cohen "Anticipation in Speech Production and its Implications for Perception" in *Proc. Symp. dynamic aspects of speech perception,* in Structure and Process in Speech Perception, A. Cohen and S. G. Nooteboom, Eds. Eindhoven, Netherlands: Springer-Verlag Berlin Heidelberg New York, 1975, pp. 124-145.

[17] A. P. Benguerel and S. Adelman, "Coarticulation of Lip Rounding and its Perception" in *Proc. Symp. dynamic aspects of speech perception,* in Structure and Process in Speech Perception, A. Cohen and S. G. Nooteboom, Eds. Eindhoven, Netherlands: Springer-Verlag Berlin Heidelberg New York, 1975, pp. 283-295.

[18] J. W. Picone, "Signal modeling techniques in speech recognition," in *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215-1247, Sept. 1993.

[19] U. Goswami, S. R. Nirmala, C. M. Vikram, S. Kalita, S. R. M. Prasanna, "Analysis of Articulation Errors in Dysarthric Speech", *J. of Psycholinguistic Res.*, vol. 49, pp. 163-174, 2020.

[20] R. Pieraccini. "from Audrey to Siri – Is Speech Recognition a Solved Problem?." The International Computer Science Institute at Berkeley. http://www.icsi.berkeley.edu/pubs/speech/audreytosiri12.pdf (Accessed Mar. 12, 2020).

[21] X. Tang, "Hybrid Hidden Markov Model and Artificial Neural Network for Automatic Speech Recognition," 2009 Pacific-Asia Conference on Circuits, Communications and Systems, Chengdu, 2009, pp. 682-685.

[22] L. Deng, G. Hinton, B. Kingsbury, "New Types of Deep Neural Network Learning for Speech Recognition and Related Applications: An Overview", IEEE Int. Conf. on Acoustics, Speech, and Signal Process., 2013.

[23] Siri. Apple.com. https://www.apple.com/ca/siri/ (Accessed Mar. 12, 2020).

[24] Amazon Alexa. "Alexa User Guide: Learn What Alexa Can Do." Amazon.com. https://www.amazon.com/b?ie=UTF8&node=17934671011 (Accessed Mar. 12, 2020).

[25] Google Assistant. Assistant.google.com. https://assistant.google.com (Accessed Mar. 12, 2020).

[26] F. Rudzicz, “Adjusting Dysarthric speech signals to be more intelligible”, *Comput. Speech and Lang.*, vol. 27, pp. 1163-1177, 2013.

[27] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanevsky, Y. Jia, “Parrotron: An End-to-End Speech-to-Speech- Conversion Models and Its Applications to Hearing-Impaired Speech and Speech Separation”, arXiv:1904.04169v3, 2019.

[28] S. O. Caballero Morales, F. Trujillo Romero “Evolutionary approach for integration of multiple pronunciation patterns for enhancement of dysarthric speech recognition” *Expert Systems with Applications*, vol. 41, no. 3, pp. 841-852, 2014.

[29] J. Wang, J. R. Green, A. Samal, Y. Yunusova, “Articulatory Distinctiveness of Vowels and Consonants: A Data-Driven Approach” *J. of Speech, Lang., and Hearing Res.,* vol., 56, pp. 1539-1551, 2013.

[30] E. Yilmaz, V. Mitra, G. Sivaraman, H. franco, “Articulatory and Bottleneck Features for Speaker-independent ASR of Dysarthric speech” *Comput. Speech and Lang.,* vol. 58, pp. 319-334, 2019.

[31] N. Terbeh, M. Labidi and M. Zrigui, "Automatic speech correction: A step to speech recognition for people with disabilities," Fourth International Conference on Information and Communication Technology and Accessibility (ICTA), Hammamet, 2013, pp. 1-6.

[32] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt, A. Hassidim, Y. Matias, “Personalizing ASR for Dysarthric and Accented Speech with Limited Data”, arXiv:1907.13511, July 2019.

[33] S. R. Shahamiri and S. S. B. Salim, "A Multi-Views Multi-Learners Approach Towards Dysarthric Speech Recognition Using Multi-Nets Artificial Neural Networks," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 5, pp. 1053-1063, Sept. 2014.

[34] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner and C. Garcia, "Dysarthric speech recognition using a convolutive bottleneck network," *2014 12th International Conference on Signal Processing (ICSP)*, Hangzhou, 2014, pp. 505-509.

[35] M. C. T A, N. Thangavelu and V. P, "Data Augmentation using virtual microphone array synthesis and multi-resolution feature extraction for isolated word dysarthric speech recognition," in *IEEE Journal of Selected Topics in Signal Processing*.

[36] Y. Takashima, T. Takiguchi and Y. Ariki, "End-to-end Dysarthric Speech Recognition Using Multiple Databases," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Brighton, United Kingdom, 2019, pp. 6395-6399.

[37] F. Xiong, J. Barker and H. Christensen, "Phonetic Analysis of Dysarthric Speech Tempo and Applications to Robust Personalised Dysarthric Speech Recognition," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 5836-5840.

[38] M. M. Sondhi, B. Gopinath, "Determination of Vocal-Tract Shape from Impulse Response at the Lips", *J. of the Acoustical Soc. of Amer.*, vol. 49, no. 6 (part 2), pp.1867-1873, June 1971.

[39] K. Richmond, S. King, P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics", *Comput. Speech and Lang.*, vol. 17, pp. 153–172, 2003.

[40] B. S. Atal, "Towards Determining Articulator Positions from the Speech Signal", in *Proc. Speech Communication Seminar*, G. Fant, Ed., Stockholm, Sweden, 1974, pp. 1-9.

[41] F. Rudzicz "Correcting Errors in Speech Recognition with Articulatory Dynamics," Univ. of Toronto.

[42] F. Rudzicz, "Production Knowledge in the Recognition of Dysarthric Speech", Ph.D. dissertation, Dept. Comput. Sci., Univ. of Toronto, Toronto, Ontario, Canada, 2011.

[43] The TORGO Database: Acoustic and articulatory speech from speakers with dysarthria, Univ. of Toronto. http://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html (Accessed Mar. 12, 2020).

[44] F. Rudzicz, "Articulatory Knowledge in the Recognition of Dysarthric Speech," *in IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 19, no. 4, pp. 947-960, May 2011.

[45] A. Wrench. "MOCHA-TIMIT." Nov. 1999. http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html (Accessed Mar. 12, 2020).

[46] Audacity® software is copyright © 1999-2019 Audacity Team. Web site: https://audacityteam.org/. It is free software distributed under the terms of the GNU General Public License. The name Audacity® is a registered trademark of Dominic Mazzoni.

[47] *MATLAB* (R2019a). The MathWorks, Inc.

[48] *Praat* (6.0.46). Praat: doing phonetics by computer.

[49] "Formant Estimation with LPC Coefficients." Mathworks.com. https://www.mathworks.com/help/signal/ug/formant-estimation-with-lpc-coefficients.html (Accessed Mar. 12, 2020).

[50] M. Yoshida, "Understanding and Teaching the Pronunciation of English." 2014. http://teachingpronunciation.weebly.com/uploads/9/5/9/1/9591739/understanding_and_teaching_the_pronunciation_of_english.pdf (Accessed Mar. 12, 2020).

[51] M. Yoshida, "The Vowels of American English." http://ocw.uci.edu/upload/files/vowels.pdf (Accessed Mar. 12, 2020).

[52] P. Ladefoged, "Vowel Contrasts" in *Vowels and Consonants – An Introduction to the Sounds of Languages,* Malden, MA, USA: Blackwell Publishers, 2001, pp. 25-30.

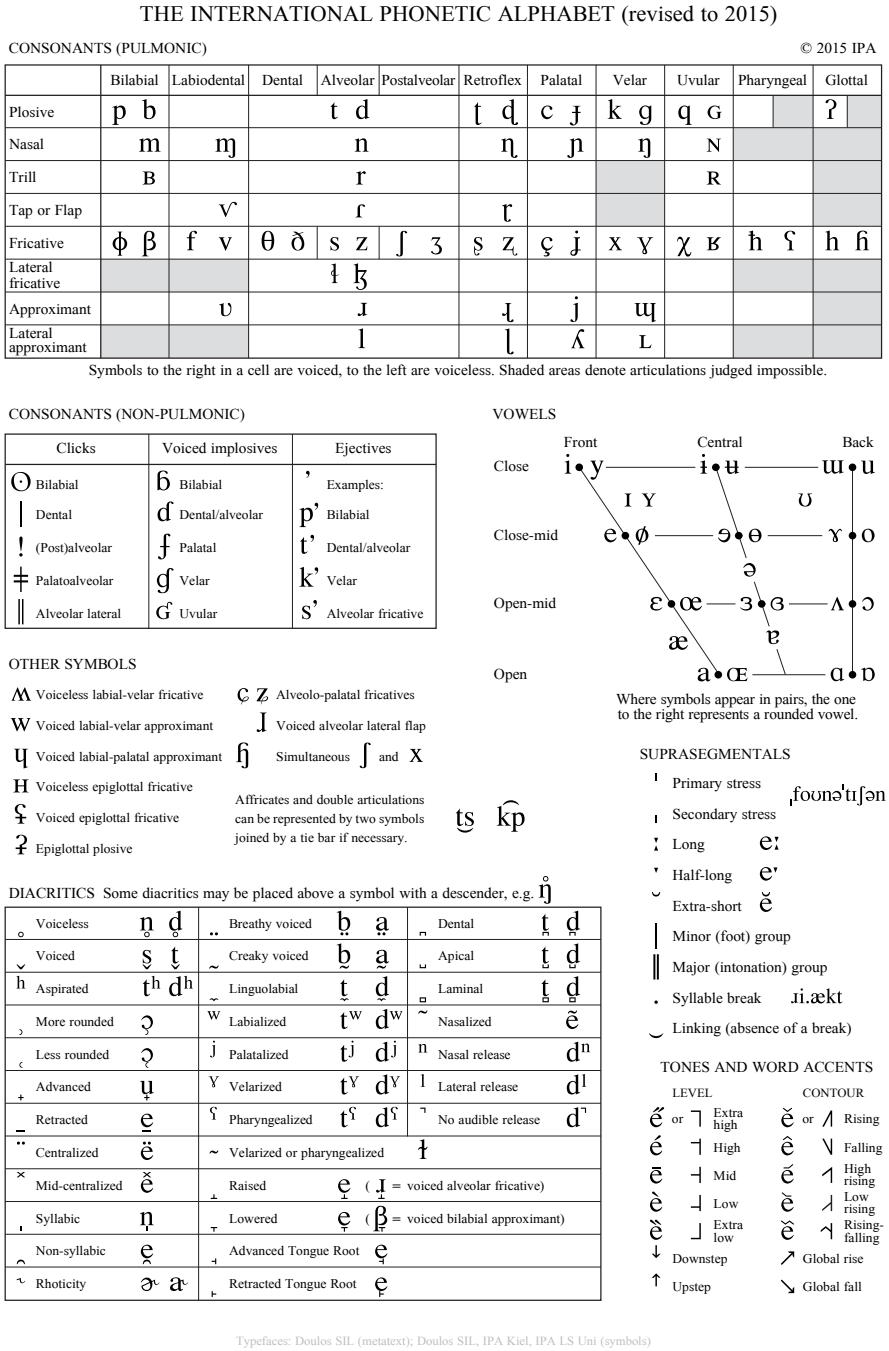
[53] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic Characteristics of American English Vowels", *J. of the Acoustical Soc. of Amer.*, vol. 97, no. 5, pp. 3099-3111, 1995.

[54] The International Phonetic Alphabet, 2015. https://www.internationalphoneticassociation.org/sites/default/files/IPA_Kiel_2015.pdf (Accessed Mar. 12, 2020).

[55] A. Klautau, "ARPABET and the TIMIT Alphabet", an archived file. https://web.archive.org/web/20160603180727/http://www.laps.ufpa.br/aldebaro/papers/ak_arpabet01.pdf (Accessed Mar. 12, 2020).

Appendix A

The International Phonetic Alphabet (IPA)



Listing A.1: The international phonetic alphabet [54].

Appendix B

ARPABET Phoneme set

Phoneme	Computer Representation		Example	Phoneme	Computer Representation		Example
	1-Character	2-Characters			1-Character	2-Characters	
i	i	IY	beat	p	p	P	pet
I	I	IH	bit	t	t	T	ten
e	e	EY	bait	k	k	K	kit
ɛ	E	EH	bet	b	b	B	bet
æ	@	AE	bat	d	d	D	debt
ɑ	a	AA	Bob	g	g	G	get
ʌ	A	AH	but	h	h	HH	hat
ɔ	c	AO	bought	f	f	F	fat
o	o	OW	boat	θ	T	TH	thing
U	U	UH	book	s	s	S	sat
u	u	UW	boot	s̆ or ʃ	S	SH	shut
ə	x	AX	about	v	v	V	vat
ɨ	X	IX	roses	ð	D	DH	that
ɝ	R	ER	bird	z	z	Z	zoo
aU or aw	W	AW	down	ž or ʒ	Z	ZH	azure
aI or ay	Y	AY	buy	c̆	C	CH	church
ɔI or ɔy	O	OY	boy	ɟ̆	J	JH	judge
y	y	Y	you	ʍ	H	WH	which
w	w	W	wit	syl l, l̩	L	EL	battle
r	r	R	rent	syl m, m̩	M	EM	bottom
l	l	L	let	syl n, n̩	N	EN	button
m	m	M	met	flapped t, ɾ	F	DX	batter
n	n	N	net	glottal stop, ʔ	Q	Q	
ŋ	G	NX	sing	Silence	–	–	
				non-speech Segment	!	!	laugh, etc.

AUXILIARY SYMBOLS (1- AND 2-CHARACTER CODES ARE IDENTICAL)

Symbol	Meaning	Symbol	Meaning
+	Morpheme boundary	:3 or .	Fall-rise or non-term juncture
/	Word boundary	* **	Comment (anything except * or **)
#	Utterance boundary	' '	Apos.-surround special symbol in comment
:	Tone group boundary	( )	Phoneme class information
:1 or .	Falling or decl. juncture	< >	Phonetic or allophonic escape
:2 or ?	Rising or inter. juncture		

STRESS REPRESENTATIONS (IF PRESENT, MUST IMMEDIATELY FOLLOW THE VOWEL)

Value	Stress Assignment	Value	Stress Assignment
0	No stress	3	Tertiary stress
1	Primary stress	ˈ	(Etc.)
2	Secondary Stress	ˌ	
		ː	

Listing B.1: ARPABET phoneme set and equal symbol from IPA phoneme set [55].

Appendix C

CMU Phonemes Set

Phoneme	Example	Translation	Class
AA	Odd	AA D	Vowel
AE	At	AE T	Vowel
AH	Hut	AA AH T	Vowel
AO	Ought	AO T	Vowel
AW	Cow	K AW	Vowel
AY	Hide	HH AY D	Vowel
B	Be	B IY	Stop
CH	Cheese	CH IY Z	Affricate
D	Dee	D IY	Stop
DH	Thee	DH IY	Fricative
EH	ED	EH D	Vowel
ER	Hurt	HH ER T	Vowel
EY	Ate	EY T	Vowel
F	Fee	F IY	Fricative
G	Green	G R IY N	Stop
HH	He	HH IY	Aspirate
IH	It	IH T	Vowel
IY	Eat	IY T	Vowel
JH	Gee	JH IY	Affricate

K	Key	K IY	Stop
L	Lee	L IY	Liquid
M	Me	M IY	Nasal
N	Knee	N IY	Nasal
NG	Ping	P IH NG	Nasal
OW	Oat	OW T	Vowel
OY	Toy	T OY	Vowel
P	Pee	P IY	Stop
R	Read	R IY D	Liquid
S	Sea	S IY	Fricative
SH	She	SH IY	Fricative
T	Tea	T IY	Stop
TH	Theta	TH EY T AH	Fricative
UH	Hood	HH UH D	Vowel
UW	Two	T UW	Vowel
V	Vee	V IY	Fricative
W	We	W IY	Semivowel
Y	Yield	Y IY L D	Semivowel
Z	Zee	Z IY	Fricative
ZH	Seizure	S IY ZH ER	Fricative

Table C.1: CMU phoneme set (After [11])