CORE
Provided by ScholarShip
Metadata, citation and similar papers at core.ac.uk

ABSTRACT

Using Synchronized Audio Mapping to Predict Velar and Pharyngeal Wall Locations during Dynamic MRI Sequences

By Pooya Rahimian

April, 2013

Director of Thesis: Dr. Nasseh Tabrizi

Major Department: Computer Science

Automatic tongue, velum (i.e., soft palate), and pharyngeal movement tracking systems provide a significant benefit for the analysis of dynamic speech movements. Studies have been conducted using ultrasound, x-ray, and Magnetic Resonance Images (MRI) to examine the dynamic nature of the articulators during speech. Simulating the movement of the tongue, velum, and pharynx is often limited by image segmentation obstacles, where, movements of the velar structures are segmented through manual tracking. These methods are extremely time-consuming, coupled with inherent noise, motion artifacts, air interfaces, and refractions often complicate the process of computer-based automatic tracking. Furthermore, image segmentation and processing techniques of velopharyngeal structures often suffer from leakage issues related to the poor image quality of the MRI and the lack of recognizable boundaries between the velum and pharynx during contact moments. Computer-based tracking algorithms are developed to overcome these disadvantages by utilizing machine learning techniques and corresponding speech signals that may be considered prior information. The purpose of this study is to illustrate a methodology to track the velum and pharynx from a MRI sequence using the Hidden Markov Model (HMM) and Mel-Frequency Cepstral Coefficients (MFCC) by analyzing the

corresponding audio signals. Auditory models such as MFCC have been widely used in Automatic Speech Recognition (ASR) systems. Our method uses customized version of the traditional approach for audio feature extraction in order to extract visual feature from the outer boundaries of the velum and the pharynx marked (selected pixel) by a novel method, The reduced audio features helps to shrink the search space of HMM and improve the system performance.

Three hundred consecutive images were tagged by the researcher. Two hundred of these images and the corresponding audio features (5 seconds) were used to train the HMM and a 2.5 second long audio file was used to test the model. The error rate was measured by calculating minimum distance between predicted and actual markers. Our model was able to track and animate dynamic articulators during the speech process in real-time with an overall accuracy of 81% considering one pixel threshold. The predicted markers (pixels) indicated the segmented structures, even though the contours of contacted areas were fuzzy and unrecognizable.

Using Synchronized Audio Mapping to Predict Velar and Pharyngeal Wall Locations during Dynamic MRI Sequences

A Thesis

Presented to the Faculty of the Department of Computer Science

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Master of Science

by

Pooya Rahimian

April, 2013

©Copyright 2013
Pooya Rahimian

Using Synchronized Audio Mapping to Predict Velar and Pharyngeal Wall Locations during Dynamic MRI Sequences

by

Pooya Rahimian

APPROVED BY:

DIRECTOR OF THESIS:____
M. H. Nassehzadeh Tabrizi,PhD

COMMITTEE MEMBER:____
Jamie Perry, PhD

COMMITTEE MEMBER:____
Masao Kishore, PhD

CHAIR OF THE DEPARTMENT OF COMPUTER SCIENCE:
____
Karl Abrahamson, PhD

DEAN OF THE GRADUATE SCHOOL:
____
Paul J. Gemperline, PhD

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Image segmentation is an integral part of computerized image analysis which consists of finding the contours of specific structural volumes or Regions of Interest (ROI) in a single image or sequence of images. ROI identifies specific sections of an image to locate the search space and identifies rough estimations of the object figures and locations. Most studies have used machine learning techniques such as Nattkemper et al. (2005), and visual features have been used to feed a supervised or unsupervised training models to predict the structural location. Image segmentation is widely applied in many fields and involves the process of identifying a structure in time (e.g., velum during elevation for a vowel) and extracting the relevant details (e.g., nasal velar surface and relationship to posterior pharyngeal wall). However, these algorithms often suffer from inextensibility and inaccuracy, as they are vulnerable in terms of the robustness in noisy environments. Dynamic Magnetic Resonance Imaging (MRI) is an imaging method that presents a great degree of noise. Thus, most powerful image processing algorithms are unable to work very well in medical image processing associated with MRIs.

Automatic tongue, velum, and pharyngeal tracking systems provide a significant benefit for the analysis of dynamic speech movements. Traditionally, movements of the velopharyngeal structures require manual tracking for image segmentation. These methods are extremely time consuming and may conceivably demonstrate inter-rater variability. Studies have been conducted using ultrasound, x-ray, and MRI to examine the dynamic nature of the articulators during speech. Noise, motion artifacts, air interfaces, and refractions often complicate the process of computer-based automatic tracking. One method to overcome the errors associated with computer-based tracking algorithms is to utilize prior knowledge to train a machine learning model. Many image segmentation algorithms work with some prior knowledge regarding the shape and/or location of target objects. In this study conventional approaches were utilized in ASR with some changes to convert the audio signal to the form of shape and figure. Speech recognition is a multidisciplinary in which many studies have been performed (Jelinek, 1998). These studies primarily have used MFCC (Ghitza, 1994; Han, Chan, Choy, & Pun, 2006; Li, Soong, & Siohan, 2000).

The purpose is to create and evaluate a model for predicting velar and pharyngeal wall positioning during dynamic MRI through the analysis of the produced audio signals. This study utilized MFCC for the audio feature extraction phase and the visual features, extracted by the researcher to feed a supervised HMM. The trained model was then used to predict the location of velar and pharyngeal structures based upon audio signals.

The approach in this study differs from other machine learning and audio feature extraction techniques; however, they possess the same fundamentals. Speech detection as the conventional in (Yu & Kobayashi, 2003) is not a concern in this study, provided that speech detection directly affects the performance of speech recognition systems. Moreover, the proposed model predicts the location of structures with a constant ratio, whereas speech recognizers detect words from a given speech signal. This approach not only works for dynamic speech analysis, but it also is compatible for problems having two main characteristics:

1. System needs time series based images (image sequence); and
2. A synchronized associated audio.

This thesis will unfold as follows:

- Chapter 2 will depict the overall picture about relevant algorithms and their applications in other fields. This chapter will also discuss why MFCC and HMM were chosen for this study.
- Chapter 3 will cover the speech production process in human and signal representation methods.
- Chapter 4 will discuss the audio feature extraction steps including pre-processing MFCC and post-processing.
- Chapter 5 will explore the proposed model and will explain the technical aspects in detail.
- Chapter 6 will present the prediction result and performance evaluation based upon two different approaches including accumulative minimum distance and inspection by researchers.
- Chapter 7 will include the conclusion and will outline possible future works.

CHAPTER 2: BACKGROUND

Image segmentation and pattern recognition have often been involved with feature extraction, boundary detection, and signal processing problems. Although these areas have been improved dramatically within the past two decades, the combination of image processing and other similar areas leads us to broaden the discussion in this chapter. Moving beyond the technical issues caused by the noise, image segmentation depends upon the content (Levine, 1969), so that the approach of segmentation may be drastically affected by the content of object(s) in the scene. Through this chapter related studies that contribute by using pattern recognition and feature extraction will be discussed, along with related works.

During the 1950's, the discussion about the differences between computers and people was a controversial topic. Although it is still an ongoing topic, even the most optimistic research could not predict the rapid growth of computers over the past decades. Polya heuristic principles (Newell, 1981), had discussed humans versus machines, in terms of performance, and his conclusion was that machines will never be able to "act" like a human. However, this idea was criticized by Greene (1959), and further studies have revealed that the Polya principle is not an accurate perception about computers, as the "meaning" behind the data plays the significant role in human and computer systems. Using the meaning behind the data leads computer systems to improve their performance, For example, Minsky (1961) used pattern recognition techniques to reduce the search space of the problem. Pattern recognition here represents a form of knowledge to improve the capability of the system by omitting inconceivable states in the search space. Recently, semantic concepts have being used to direct the search (Tang, Xu, & Dwarkadas, 2003) by reducing the search space. Although the application of knowledge may be limited to some specific problems, the general idea of using knowledge is rapidly extended in many areas like medical image processing (Clark et al., 1998). However, the crucial question is "How we extract knowledge from data?" and "How we apply the knowledge to solve the problem?"

To answer these questions, the difference between prior knowledge and feature extraction must first be distinguished. When people look at images, prior knowledge helps them to understand the concept or the meaning of the image. This exploration may be conducted by the segmentation of elements and objects on the scene. Segmentation utilizes the prior knowledge as well as the context of the image. For example, when a person looks at an island sunset; some objects, such as a microwave, may not be appropriate to be in the picture frame. Thus, the context of image restricts the availability of objects. Moreover, pre-knowledge about the shape, color, and figure of objects may lead the observer to segment the image (Comaniciu & Meer, 1997; Cremers, Rousson, & Deriche, 2007). These forms of pre-knowledge may be applicable for the human observer; Yet, in terms of computer segmentation, there is another issue that is introduced where the computer system needs some indicators, called "features," from the original data whether spatially or temporally, where eventually they represent the characteristic of the data and make them unique and comparable. Various features extraction has been introduced in a wide spectrum of computer areas. For example, the application of feature extraction has been used for the prediction of the stock price index (Kim & Han, 2000) or algebraic feature extraction in image recognition problems (Hong, 1991). The features may be driven selectively based upon the content of problems; Thus, the deep understanding of problem characteristics would be a vital step.

Guyon et al. (2006), introduced the different dimensions of feature extraction and feature selection principles in the book *Feature Extraction: Foundations and Applications* (Guyon, Gunn, Nikravesh, & Zadeh, 2006), that would be helpful to be briefly explained here. The feature extraction pre-processing steps are explained by introducing some notations: “Let *x* be a pattern vector of dimension *n*, $x = [x_1, x_2... x_n]$. The components x_1 of this vector are the original features. *x* is a vector of transformed features of dimension *n*” (Guyon et al., 2006) the pre-process transformation would be the following steps:

Standardization: Regardless of the content of patterns and their calculations, they have to be in the same scale and format. For example, temperature may be a component in Celsius and the other component would be in Fahrenheit. Several mathematical operators would be applicable for both of them but they may not be meaningful before being transformed into a unified format.

Normalization: The image contrast would be a good example to introduce the normalization pre-processing. In medical image processing, different wave absorption causes different image contrasts. So, the normalization pre-process would be the conversion image contrast into a specific range.

Signal enhancement: One of the most important factors is the signal-to noise ratio which may be improved by applying filters and smoother algorithms.

Extraction of local features: There are many different algorithms that have been invented about local feature extraction including synthetic and structural methods.

Linear and non-linear space embedding methods: In the real world, a large dimension of data is involved. Data reduction usually contributes to the loss of useful information, yet the system is unable to work with a large scale of data; Consequently, the dimension of data has to be reduced to a reasonable volume, with a minimal amount of lost information. Principal

Component Analysis (PCA: Jolliffe, 1986), and Multidimensional Scaling (MDS: Kruskal, 1964), are examples of space embedding methods.

Non-Linear expansions: Due to the problem complexity, the feature dimensions have to be extended. This process is the reversed version of previous item. In these cases, the direct features extracted are not sufficient for a proper training; Therefore, the features have to be expanded by some indirect methods.

Feature discretization: Basically, continuous data is not well-formatted for many algorithms; Thus, according to the designated learning machine, data may need to be discretized. For example HMM originally works with discrete data, provided that the Gaussian Mixture Model (GMM) works with continuous data.

Guyon et al., (2006) believes that feature selection may involve data reduction in the feature set, the main data, and the performance consideration. Initially, to illustrate how much data reductions have had an adverse influence on the main data, a mechanism is required to visualize features and main data to evaluate the effect of data reduction and the probable loss of useful data.

In this sense, Nilsson (1969) introduces three feature extraction facts, consisting of the following: 1) There is no straight-forward instruction or algorithm to extract features and it deeply depends upon the nature of problems; 2) Designing a feature extraction is a practical process involving different policies and considerations; and 3) Biological prototypes such as humans would inspire researchers to invent a new feature extraction approach. Consequently, finding attributes that describe the differences and characteristics of the source may be variant for the given problems and constraints. However, this challenge is an integral part of feature extraction, and some studies split it into microanalysis and macro-analysis referring to local and global processing respectively (Brice & Fennema, 1970). Underneath all of this categorization, several subcategories have been introduced such as smoothing, line operation, contour, edge detection, curvature detection, and normal representation. Handwriting recognition is the best example to find the application of those feature extraction methods in order to recognize isolated characters. The feature extraction is the most important section of a handwriting recognizer. Due Trier et al. (1996) reported a wide spectrum of feature extraction methods in handwriting problem (Due Trier, Jain, & Taxt, 1996). Optical Character Recognition (OCR) was introduced and then the topic was diverged on feature extraction methods that have been contributed to improve OCR performance. Although the proposed algorithms revolve around the pattern recognition in image processing, feature extraction is a broad concept beyond of those topics.

For instance, audio feature extraction is a well-established area and the result of several successful studies are available as on-shelf products, for example Apple[1] iOS 6 has been enhanced by a powerful speech recognizer system and one of the most important parts of its units is audio feature extraction, since the audio signal basically is meaningless for the system. Therefore, some specifications have to be extracted from the audio signal as a set of features. These audio features represent those properties of audio signal to enable a classifier to determine utterance words.

ASR is found on the top of feature extraction concretes to convert human speech into a sequence of text in real-time. However, spoken words may be the same for different speakers, and in terms of signal analysis they may be completely different. Even one speaker may generate different form of signals for one specific word. So the system should recognize patterns of words

[1] Apple Company: http://www.apple.com/ios/siri/

by analyzing the signal. Some ASR systems have been adapted from human auditory models such as perceptual linear predictive (Hermansky, 1990). In this sense, the human auditory system is closely explored and some characteristics of the model are adapted to design a new computer model. MFCC is another example of audio feature extraction that will be discussed in this chapter and in chapter 4 in detail. There is a mutual collaboration between human auditory system studies and ASR systems and thereby the influence of ASR systems on human auditory is noticeable and vice versa. Aside from feature extraction based upon the human auditory model, the ASR advanced methods have been utilized as tools in speech therapy or audio perception studies such as usage of ASR for voice therapy assessment (Kitzing, Maier, & Åhlander, 2009), the application of ASR in pronunciation training (Dalby & Kewley-Port, 1999), and a proper replacement for human expert to evaluate the speech by using ASR systems (Riedhammer et al., 2007). The significant influence of audio feature extraction in speech therapy is rapidly increasing as a result of the invention of more accurate computer data acquisition systems. On the other side of this collaboration, there is a bulk of studies that have been inspired by the human auditory model. For example, Holmberg (2006), introduced a simplified model based upon the human auditory model and then compared the proposed model with other resemblance models such as PLP and MFCC (Holmberg, Gelbart, & Hemmert, 2006).

Although ASR systems are well-established, they still have some issues with images in noisy environments. Consequently, pre-knowledge information may reduce the effect of noise and may improve the performance. This pre-knowledge information may be found in different forms, but this chapter specifically discusses various methods in audio-visual fusion systems, as this approach is close to the proposed solution in this thesis. The book, “Hearing by Eye Two” fundamentally explores the models of audio-visual fusion systems (Campbell & Dodd, 1980).

The basic idea behind of audio-visual systems is applying visual features to improve the machine learning performance in presence of noise and distractions. Audio-visual speech recognition looks like human hearing mechanism in a crowded place. We usually concentrate on lips of speaker when we are in a noisy place and we use visual information to improve the hearing result by lips reading. Campbell and Dodd (1980) introduced four different models of audio-visual integration including direct identification model, separate identification, dominant recording model and motor recording model. Through the direct identification model which is adapted from Klatt (1979) research, the facial specifications feed the classifier. Separate identification models have to be coded separately for a specific phonemic feature and then it uses a fusion model to find the phonetics. The Dominant recording model introduces an audio signal as the dominant modality and then visual features are extracted separately and fused. Prediction is merely based upon an/the auditory model. The Motor recording model is based upon Campbell's categorization "both input are projected into an amodal (neither auditory nor visual) common space and fused in the space" (Campbell & Dodd, 1980); the visual features based upon some of vocal cavity structures configurations are classified and the system introduces a mapping between words and figures. This model has to consider many different combinations of structures provided that some of them are visible (e.g. lips) and that some of them are not visible, such as the velum and certain factors, such as velocity and trajectory, also play the significant role in these systems. Zeng (2009), illustrates an audio-visual fusion expansion method in behavioral science and psychology studies to investigate on human behaviors (Zeng, Pantic, Roisman, & Huang, 2009). As previously discussed, audio-visual speech recognitions are ASR systems that are tuned by a set of visual features as a result of better performance in the presence of noise, because using visual features is a way to compensate for the adverse influence of the noise on the extracted audio features. Although MFCC as audio feature extraction has been used in ASR and audio-visual speech recognition systems, it has been a widely accepted algorithm in music analysis. To classify different genres of music, there are two approaches that have been introduced, consisting of standard feature extraction methods as the classical approach, such as Zero-crossing, and considering the human auditory perception, such as MFCC. McKinney & Breebaart (2003), reported the performance of both categories of feature extractions. In this study, low-level signal properties, MFCC, psychoacoustic features have been discussed and they introduced a new method called "auditory filter-bank temporal envelopers" (AFTE) were evaluated in five different music genres. The authors extended the evaluation by dividing Popular Music into seven different sub-groups. The classifiers were the same for all feature extractors. However, the performance of the classifiers depends upon the genre of music, and this study reveals that the AFTE has better performance overall. This study is an example of MFCC usage in music analysis. In terms of music analysis, Logan claims that MFCC does not affect the prediction adversely (Logan, 2000). Also MFCC may be used in musical instrument recognition (Krishna & Sreenivas, 2004), or finding similar music based upon audio signal analysis (Logan & Salomon, 2001). Consequently, MFCC has been recognized as the dominant audio feature extraction method in both human voice and music fields. Another piece of this puzzle is classifiers, which is an extensive topic, as classification is not limited to audio signals or image classification.

Machine learning problems involve a series of input to train the model and to notice the correlation between input patterns and the desired output (Guyon et al., 2006). The model represents the association between the input pattern and the output which is called "learning machine" where basically it is a function to mimic the system behavior by optimal input dimension. The “learning process” is divided into three categories based upon how the system is trained. The classes may be predefined in *supervised* models; in contrast, classes are unlabeled in *unsupervised* models. There is the third category of learning which is called *reinforcement* learning. In this learning model, a function evaluates the accuracy of the classifier to illustrate how the system works accurately without any further information. The system has to improve the performance by analyzing the function feedback gradually through the process. Supervised classification is one of the most famous classification categories that have been conducted in intelligent systems (Kotsiantis, Zaharakis, & Pintelas, 2007). Decision trees are the simplest form of classification by representing possible classes as the leaves of a tree and accordingly, each branch may have a different weight. The crucial point of a decision tree is finding the best root (starting point), as finding the proper root may contribute to the improvement of the performance of the classifier. The decision tree may be converted to a form of Disjunctive Normal Form (DNF) rules wherein some problems, it would be a suitable solution with a reasonable error rate and speed. They are appropriate methods for “single feature at each internal node” (Kotsiantis et al., 2007). Another category of learning machines is working based upon the sum of perceptions over the graph. If the result was more than the threshold the output would be 1; otherwise, it would be 0. The trained model will find the best label for the input sequence during the testing phase. Through the linear classifiers, the purpose is to find a line to separate different instances. In the case of two dimensional inputs, a plane would separate classes; However, this method is unable to satisfy more than two dimensional inputs. Therefore, a nonlinear approach would be the solution for instance; the Artificial Neural Network (ANN) would be introduced to overcome this problem. In a multi-layer ANN, input data passes through the input layer, a set of neurons connecting to the hidden layer of network, and the output would be addressed by the output layer receiving data from hidden layers. The combination of hidden layers is variant but the input should be the same as the input vector dimension, and the output layer has to be the same as the output vector dimension (Hagan, Demuth, & Beale, 1996). The weights of the net may be adjusted by the training set. Super Vector Machines (SVM) is another classifier that solves the problem of "maximizing the margin and largest possible distance between the separate hyperplanes" (Kotsiantis et al., 2007). A SVM is a proper learning mechanism for larger number of features. The performance of SVM is not affected by input dimension; Consequently, this method is a very suitable approach for complicated and nonlinear problems. SVM may suffer from misclassified instances contributing to misclassification or in some cases; It may not be able to introduce a solid separator for those overlapped classes (Konstantinos Veropoulos, Campbell, & Cristianini, 1999). SVM have been widely used in tissue image segmentation and cancer diagnosis; for example, Cataldo et al, proposed a cancer cell segmentation method based upon a supervised SVM classifier. With a breast cancer diagnosis, Geraldo et al. used several measurements, such as Geary's coefficient and Moran's index (Moran, 1950), to train a SVM classifier, and then they measured the accuracy of the system (Braz Junior, Cardoso de Paiva, Corrêa Silva, & Cesar Muniz de Oliveira, 2009). There are several studies such as (Junior, Paiva, Silva, & de Oliveira, 2009), (Nunes, Silva, & Paiva, 2010), (Wang, Zhu, & Liang, 2001), and (K Veropoulos, Cristianini, & Campbell, 1999), that reveal that SVM is a pivot point of many studies revolving around diagnosis classification problems. The last classification approach that needs to be discussed is HMM. Basically, HMM may be continuous or discrete classifiers, but there are several other versions that have been introduced that work with continuous data with different architecture such as the GMM. HMM, the same as other models is a powerful classifier that has been applied in a variety of studies. One of the most dominant applications of HMM is when it is used in ASR systems. Aside from speech recognition, HMM is a proper model in brain-computer interfacing such as (Obermaier, Guger, Neuper, & Pfurtscheller, 2001), and (Chiappa & Bengio, 2004) studies.

One question may arise: Why was HMM was designated for this thesis? Some reports have repeatedly proven that HMM performance is better than SVM such as in (Yi-Lin & Gang, 2005). Yi-Lin et, al. classified five emotions by both HMM and SVM. They achieved 98.9% accuracy for female subjects and 100% accuracy for male subjects in terms of prediction emotion based merely upon audio signals and using HMM, providing that SVM was less accurate. In some studies, such as (Jianjun, Hongxun, & Feng, 2004), a hybrid SMV/HMM improves the performance by compensating for their drawbacks. The hybrid model has been evaluated on various standard data sets and the result shows that the hybrid model performed better than the GMM; However, the difference was not significant. Moreover, in this proposed study audio signal as pre-knowledge will be used to convert audio features into a set of visual features where these features may be used as ROI for other image segmentation methods, or they may be independently applied in tracking systems. Consequently, HMM would be the best choice for the model. Besides that, HMM is the appropriate model for temporal problems, considering that the analysis of audio signal and video are temporal topics. For visual feature extraction, the features were extracted by the researcher manually. The proposed model combined the audio-visual features to train the model in order to predict the location of velum and pharyngeal wall based upon the audio input signal. In chapters 4 and 5, the internal mechanism of HMM and changes made to ASR systems for compatibility purposes will be discussed. In addition, the visual feature extraction mechanism will be discussed in the following chapters but before that, the human auditory model as the initial point and the target of this study, has to briefly be reviewed.

CHAPTER 3: SPEECH PRODUCTION AND SIGNAL REPRESENTATION

3.1 The Respiratory System and Sound Production

“Speech,” in technical terms, refers to the acoustic presentation of language (articulation, fluency and voice); but moving beyond the language concepts, speech is a mechanism by which the respiratory organs perform as the provider of speech (Forney Jr, 1973), which creates air pressure and the mechanical process of vibrating the outgoing air that generates sounds. Air is supplied by the lungs surrounded by the rib cage and the diaphragm muscle. The rib cage rises when the diaphragm muscle contracts (Sharp, Goldberg, Druz, & Danon, 1975). Therefore, the size of the rib cage is increased and the air flows to the lungs, which is called *inhalation*. *Exhalation* is the reverse process of inhalation, by shrinking the size of the rib cage and respectively, the lungs (Nilsson & Ejnarsson, 2002). However, the respiratory process is a semi-involuntary mechanism and the duration of exhalation and inhalation in breathing are the same; the duration of exhalation may be different based upon the situation (i.e. speaking).

3.1.1 Vocal Fold

The vocal tract is the channel that mainly consists of the *phonatory system* (or larynx) which is connected to the trachea, and moves up to the nasal cavity. The larynx prevents food and liquids from aspirating into the lungs and additionally, it plays a significant role in speech production. The vocal cord is an opening with two horizontal membranes across the larynx opening. During exhalation, the airflow coming out from larynx is manipulated by the vibration of these muscles. However, the vocal cord produces a *buzzing sound*, and other organs (i.e. Lips, tongue, jaw, velum, and pharyngeal wall) in the vocal tract modulate the sound to produce variant sounds (Peterson & Barney, 1952). The vocal fold has three different states: voiced, voiceless, and breathing states. During the breathing state, the vocal fold is open and both of its muscles are relaxed, allowing the vocal fold's muscles to block the airflow path partially during the voice production, allows airflow to pass through the glottis (Nilsson & Ejnarsson, 2002).

The surrounding area among the lips, teeth, velum, and larynx are called the *"vocal tract,"* as shown in Figure 3.1, and it may be merged with the *nasal cavity* through the velum. The output frequency may be manipulated by various configurations of the tongue, lips, and velum. Even the shape, movement, and velocity of tongue and velum dramatically affect the output frequency. The velum is a biological valve that has an important role in speech production by controlling the connection of the nasal cavity and vocal tract. The velum merges the nasal cavity and vocal tract when necessary and it excludes the nasal cavity from the vocal tract to pass air flow to the vocal cavity. As a result of the smaller size of the vocal folds and the higher frequency of vibration, female and children generally have higher pitched voices (Nilsson & Ejnarsson, 2002).

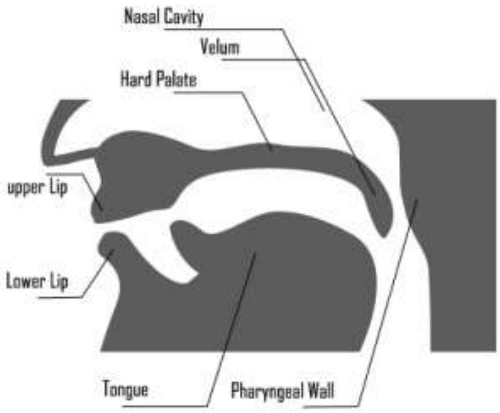


Figure 3.1 Human Vocal System

The lungs' pressure is increased by the diaphragm, and air flows from the lungs to the vocal folds. When the air pressure decreases behind the vocal cords, as a result of the elasticity of the fold and Bernoulli Effect, muscles move back into place (Bellman, 1956). In contrast, the difference in pressure of both sides of the vocal cords forces muscles to be opened and therefore, air flows up to the vocal cavity. Due to the air pressure reduction, vocal cords smoothly revert back to the normal position. The frequency of vibration (rushed opening and closing) in women is generally more than that in men, and it is another reason for the higher pitch in most women.

Figure 3.2 accounts for the discrete-time speech production model. The Impulse Generator and Random Noise Generator in this model play the role of the excitation generator or lungs in the biological version (Benesty, 2008). The impulse generator excites the glottal filter (e.g. vocal cords) (Howard, 1960). $u[n]$ is the input of the vocal tract and G is the gain of voice volume. A time-varying Digital Box may be used with different filters based upon the model. For example, according to the human anatomy, the vocal tract and lips/jaw filters may be considered filters that manipulate output signals and finally, $S[n]$ is the output speech signal of the model.

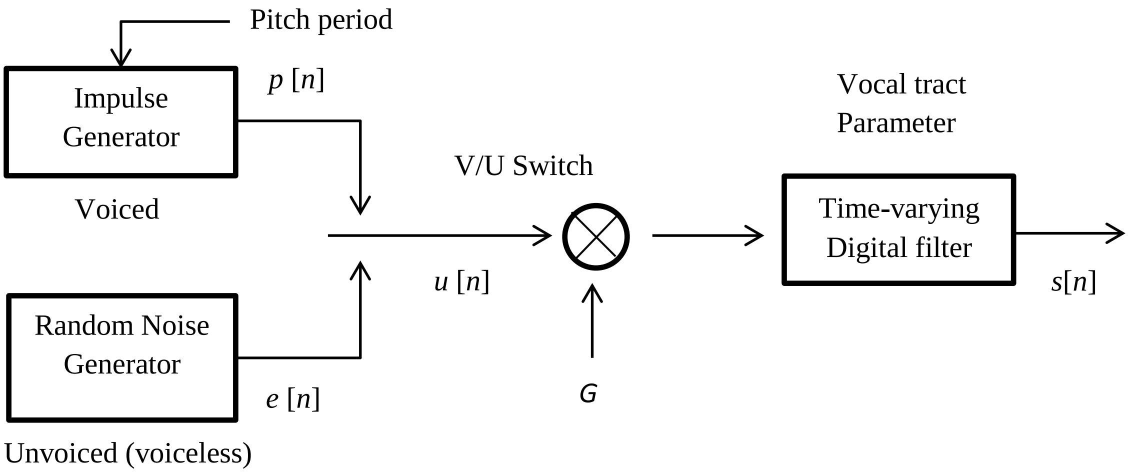


Figure 3.2 Discrete-Time Speech Production Model (Benesty, 2008)

3.2 Signal Representation

There are two major signal representations based upon time or frequency domain. In computer speech processing, however, three various states may be considered: voiced, voiceless, and silence. Although, there is no clear boundary among these states, there have been several studies in this area to recognize voice boundary from silence and voiceless states (Bachu, Kopparthi, Adapa, & Barkana; Qi & Bao, 2006; Radmard, Hadavi, Ghaemmaghami, & Nayebi, 2011). In the silence state, there is no speech signal. Voiced sound is the representation of the vocal cords' oscillatory vibration and its modeled unvoiced sounds are very similar to noises (Benesty, 2008).

As shown in Figure 3.3, a sample of a speech signal was selected to demonstrate the three different states of a speech signal. The first selected section indicates silence mode; the next is unvoiced, as shown in the larger plot, and this section accounts for random noise. The last marked section is voiced with a high amplitude value. The combination of these three types of produced signals introduces different types of vocal presentations, such as a whisper.

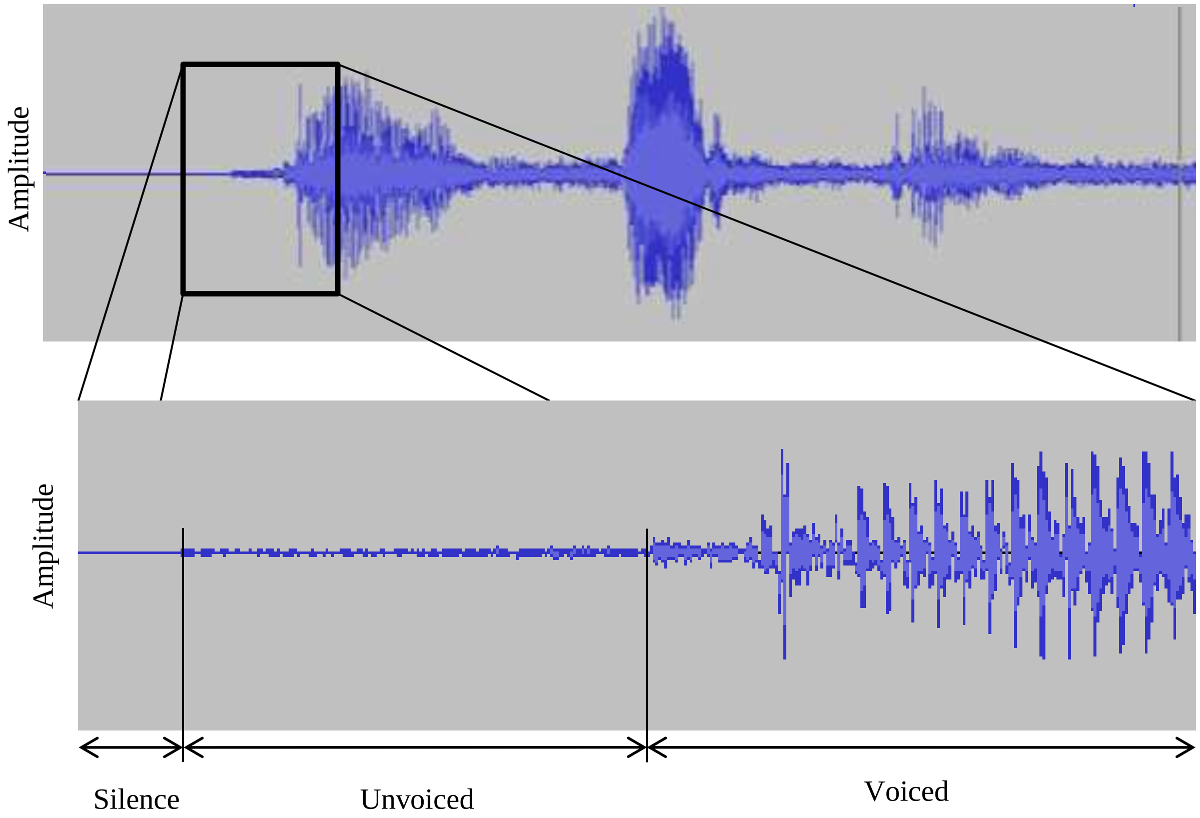


Figure 3.3 Silence, Unvoiced and Voiced Signals

Given the speech signal in Figure 3.4, the Spectrogram is the spectral representation of sound frequency in Figure 3.5. In order to calculate the spectrogram in time and frequency domain in this study; as shown in Table 3.1, *nwin* is the Hamming Windowing length and it is the same as *nfft* (Fast Fourier Transform length). *fs* variable is the sampling frequency of input signal. The *Noverlap* variable shows the segment overlapping where it is at 50% of *nwin* variable.

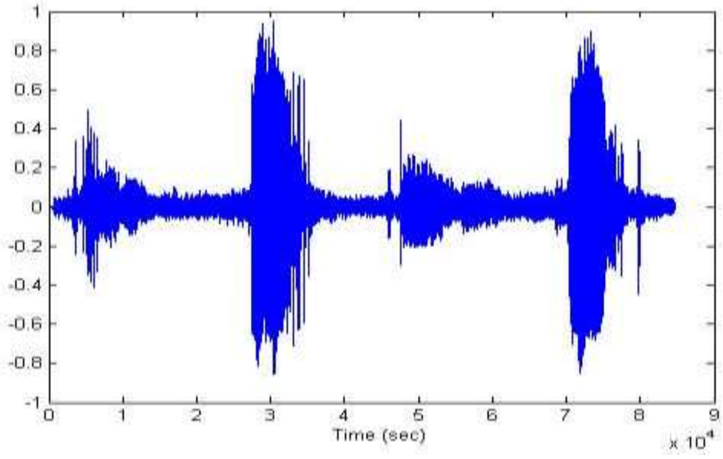


Figure 3.4 Speech Ware

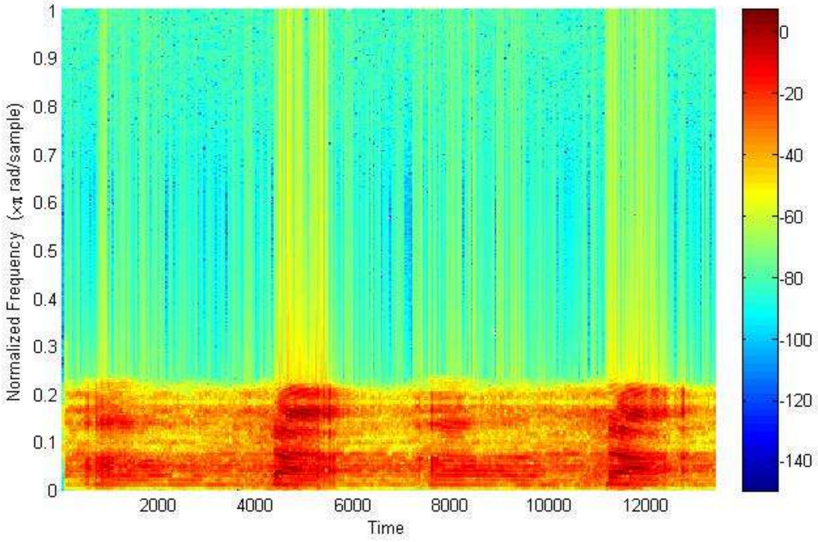


Figure 3.5 Corresponding Spectrogram

```
[y, fs]=wavread(‘sound.wav');
nwin = 512;
noverlap = 256;
nfft = 512;
spectrogram(y, nwin, noverlap, nfft,  'yaxis');
colorbar
```

Table 3.1 Spectrogram MATLAB Code

3.3 Chapter Summary

There are several studies that have been published involving the physiological aspect of speech structures. The lungs, as the supplier of airflow for producing speech play a crucial role; Moreover, the vocal cords vibrate when airflow exits through the vocal cords and it produces a voice. The vocal tract generates sound by deforming the tongue, lips, and velum. The nasal cavity adds another tube to lead airflow through the nasal and oral cavities.

Most studies about the vocal tract explore the dynamic movement of the articulatory structure of the oral cavity, since there are no moving organs within the nasal cavity; that said, in the oral cavity, (vocal tract) parts are movable and their composition produces different sounds. Thus, studies on the vocal tract are time consuming, and they require the analysis of signal production and the trajectory of structures.

This chapter has covered the basic process of voice production from the physiological standpoint. This study will focus on the velum and pharyngeal wall movement prediction during speech production.

CHAPTER 4: AUDIO FEATURE EXTRACTION AND HIDDEN MARKOV MODEL

In this chapter, some well-known methods and algorithms are adapted from ASR systems, mainly audio feature extraction algorithms and the HMM. Based upon the nature of this study, certain steps were changed in the audio feature extraction algorithm in order to achieve the proper audio feature vectors synchronizing with the visual features. In order to design and train a model, the HMM will be described in this chapter with an example describing a prediction based upon observations. Along with the example, the Forward-Backward algorithm will be explored since it is the foundation of design, training, and prediction phases.

The HMM will be described and three issues in this model will be discussed. The HMM will be used in a wide spectrum of problems associated with prediction and learning. This chapter is an introduction of this model.

Although feature extraction is the foundation of this stud and ASR systems, the output of ASR and this study is completely different showing in Table 4.1. Regardless of utterance length, ASR produces corresponding words provided that in this proposed model the output length is a coefficient of utterance length.

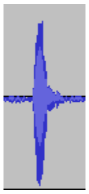
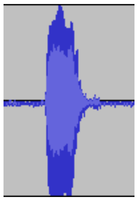
Signal	**Automatic Speech Recognition system**	**The proposed system's output**
 Duration = *k/2*	"Book"	$\begin{pmatrix} x_{1,1} & \cdots & x_{1,n/2} \\ \vdots & \ddots & \vdots \\ y_{m,1} & \cdots & y_{m,n/2} \end{pmatrix}$
 Duration = *k*	"Book"	$\begin{pmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ y_{m,1} & \cdots & y_{m.n} \end{pmatrix}$

Table 4.1 the Proposed Model Output and ASR System Output

There are two differences between the proposed model and ASR systems:

1. Voice Activation Detection (VAD) plays a significant role in the ASR system in order to segment each "word" in the signal stream, provided that in this study, there is no "word" at this point and the system should conceivably predict the

figure and location of structures, regardless of whether that speaker is speaking or not.

2. In ASR Systems, the HMM absorbs the effect of the various lengths of signals representing words (Table 4.1), but in this study, 40 predicted locations per second are generated without considering utterance states (voiced/ voiceless/ silence).

However, while both of these systems use the same tools and algorithms, there is a significant gap between these studies in terms of generating more accurate results.

There are many audio feature extraction methods that have been developed based upon the type of input signals. Some of the features may work properly with certain specific training models while some do not. Basically, the audio feature extraction is achieved by selecting some specifications and properties of audio signals. These extracted features demonstrate one or more than one characteristic of the input signal, and depending upon the utterance/signal generator, different audio feature extraction methods may be applicable. Still, choosing the ideal one is always a challenging task. The majority of audio feature extraction work is based upon linear coded signals and most well-known methods are adapted from this approach. MFCC, one of the most dominant audio feature extraction methods, is categorized in cepstral domain where features are calculated in a short time (steady feature) between 10-30 ms in length. Most audio feature extraction methods are the sequence of components transferring the features from one domain to another domain (Liu, Wang, & Chen, 1998). The design of a new method is a complicated process, since the signal characteristic plays the crucial role in the design of such methodology. Moreover, transferring from one domain to another may potentially affect the interpretation of features.

4.1 Audio Feature Extraction

Temporal domain is a time based domain that considers waveform changes through time, and is the base of feature extraction methods and in later steps, is converted to other domains as shown in Figure 4.1. However, feature extraction is about the quantifying audio signals to some vectors, as it always has been involved with the noise issue. To achieve noise removal, certain pre steps must be performed before the feature extraction state. Cleaning the signal and noise removal are the main purposes of the pre-processing, and then the filtered signals must pass through the Frame Blocking, Windowing, FFT, Mel-Frequency Wrapping, and Cepstrum, and finally, post-processing is performed, preparing and readying the feature vectors.

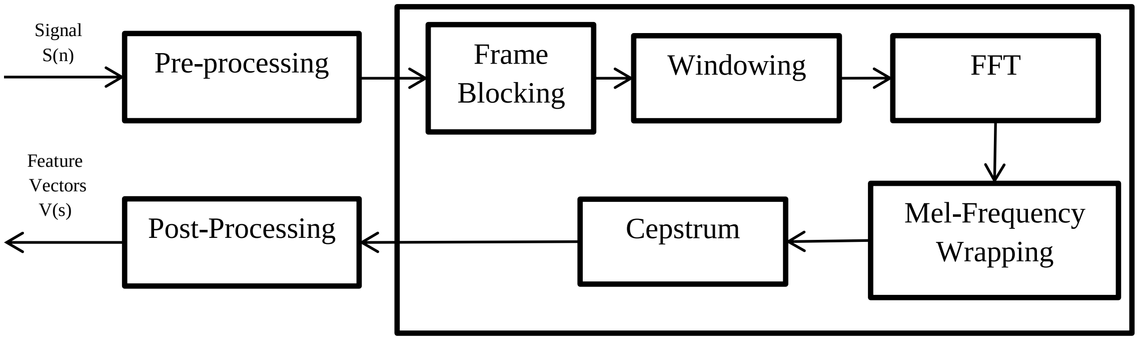


Figure 4.1 Audio Feature Extractions

4.2 Preprocessing

This step prepares the audio signal for feature extraction through changing sampling rate, noise removal, and Pre-emphasis blocks in the pre-processing chain. The previously mentioned block sequence occurs with VAD in order to determine the endpoint of utterance. Determination of utterance speech has been a big challenge for speech recognizers, because incorrect endpoints may contribute to the dramatic reduction of the recognizer performance. To overcome this issue, Short-Term Energy estimation, Short Term Power, and Zero-Crossing are generally applied. This block was not applicable in this study; however, speech recognizers need to detect utterance speech duration in order to determine the best-fit word for the specific chunk input signal, provided that the proposed model will predict certain markers over the boundary of the velum and pharyngeal wall regardless of whether the speaker is talking or not. In other words, the work-unit of the proposed system is “time,” while speech recognizers accept signals and then split the input stream to chunk of utterance speech; finally, they determine the most possible corresponding ”words.” Thus, Figure 4.2 demonstrates the pre-processing block for the system:

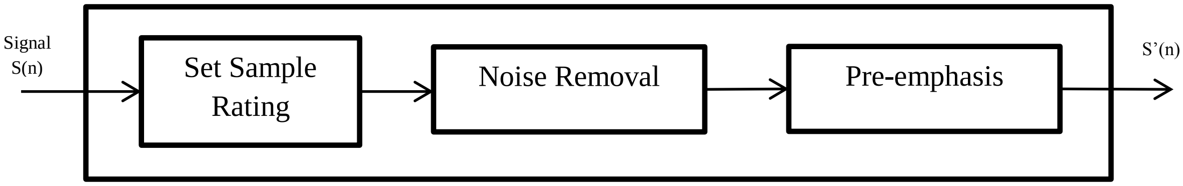


Figure 4.2 Proposed Pre-processing Blocks

4.2.1 Sample Rating

There are several studies that have been done about the impact of sample rating on training models (Ssnderson & Paliwal, 1997). Two important factors impact training models involving the length of vectors and the sampling rate. However, both factors affect training, and the type of training model may be the third factor. With respect to the MFCC, the ideal range of sampling rate is between 14K and 16K. In this study, a 16K sampling rate was designated for the feature extraction phase. As a result of the changing of the sample rate of the input signal, the

Audacity[2] software, an open-source application under the GNU General Public License, was used.

After opening the audio file in Audacity, it should be converted to a mono signal. Audacity implements the conversion by selecting *tracks* > *Stereo Track to Mono*. After running the command from the menu, both left and right signals are merged into the mono format. For the next step, the sampling rate should be changed to 16,000Hz. The mono signal has to be converted to the 16K Hz by Selecting *tracks* > *Resampling* > *16000* Hz.

4.2.2 Noise Removal

Fourier analysis is the core algorithm for Audacity noise removal. The noise removal phase has two internal steps including *picking the noise segment* and *noise removal* based upon a threshold. In the first step, a chunk of noise (silence section) should be segmented as the frequency spectrum of the noise; then, selected background noise is compared with the input signal and any tones that are not louder than the segmented would be reduced in terms of volume.

4.2.3 Pre-emphasis Digital Filter

The pre-emphasis digital filter improves the signal quality in order to minimize the noise ratio by reducing the difference between high and low amplitude frequencies. Generally, Eq.4.1consists of two parts:

[2] http://audacity.sourceforge.net/

$$Y(n) = X(n) - a.X(n-1) \tag{4.1}$$

Where $X(n)$ is the input signal at n and a is a constant (between 0.95 and 0.98), in the finite impulse response (Saramäki, 1993). Eq.4.2 and Eq.4.3 formulas represent Finite Impulse Response (FIR) filter:

$$H(z) = 1 - 0.95z^{-1} \tag{4.2}$$

$$\mathrm{y[n]} = \sum_{k=0}^{n} h_k \, s[n-k] \tag{4.3}$$

In this study, 0.95 was designated to be the constant (a) of the filter.

4.3 Frequency Cepstral Confidents

The audio feature extraction techniques are widely expanded in many areas associated with audio signals and classification problems. For example, Short Time Energy and Zero-Crossing Rate are widely used in the short-time analysis of speech signals, music, and silence recognition. Short-Time Energy represents the amplitude variation over time. Energy function values in a voiced speech signal are significantly higher than in an unvoiced speech signal making it a remarkable feature extraction method for voice recognition problems. MFCC works more accurately in human voice frequency (Ghitza, 1994). MFCC, one of the most well-known audio feature extraction methods as was discussed in chapter 2, has been widely used in speech recognition studies. The Mel-Frequency Scale reflects the human ear perception and imitates the hearing process of humans by utilizing the Mel-Warped Frequency Scale. Low frequency resolution plays a significant role in the formants capturing process. Close spaced overlapping triangles in low frequency region of MFCC and fewer numbers of triangular filters in the high frequency region create a better resolution in low frequency.

4.3.1 Frame Blocking

Frame blocking selects the frame size and based on *m* samples separation. In this study, for a 16KHz sampling rate, 20 ms was assumed stationary and consequently, and therefore, the frame length would be 320 samples. Overlapping decreases the separation factor to 120 samples by considering 62.5% overlap rate.

As show in Figure 4.3, *m* is the length of the new part of the block and additionally, *length-m* is the overlap of the block; therefore the step of blocking is *m.*

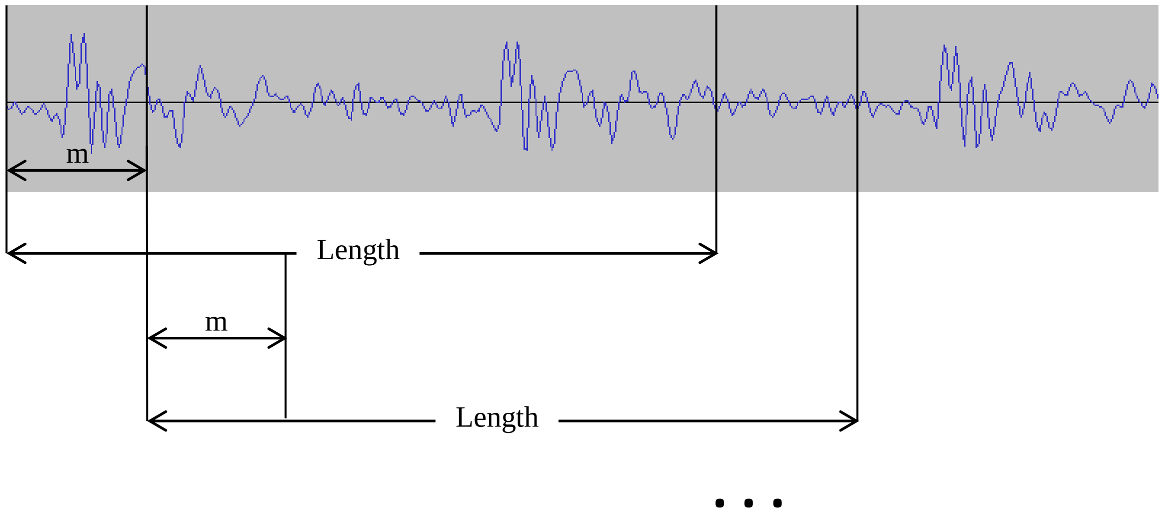


Figure 4.3 Frame Blocking

4.3.2 Hamming Windowing

Hamming Windowing is used to cancel the effect of signal discontinuity at both sides of blocks. The corresponding Eq.4.4 for the Hamming Window calculation is:

$$\mathrm{W(k)} = 0.54 - 0.46\ \cos\left(\frac{2\pi k}{k-1}\right) \tag{4.4}$$

Hamming Window is a bell-shape in which the vertical axis demonstrates amplitude and the horizontal axis accounts for number of samples in each block (*length*). The result of applying the Hamming Window is that the beginning and end point of each block gradually are converged to zero, thereby reducing the adverse consequence of signal discontinuity effect.

4.3.3 Fast Fourier Transform and Feature Extraction

In this phase, the corresponding features should be extracted from the blocks. There are a myriad of algorithms that have been invented to extract features from audio signals such as linear prediction (Makhoul, 1975) and Mel-cepstrum methods. Many studies have revealed that Mel-cepstrum has a better performance than other methods (Qingzhong, Sung, & Mengyu, 2009), (Pearce, 2000), because this approach is able to mimic the human auditory model better than other approaches. With the Mel-cepstrum feature extraction, the next step would involve the calculation of the Fast Fourier transformation (FFT) for each block and then calculated the Mel-Wrapping and then the corresponding cepstrum.

4.3.4 Fast Fourier Transform and Mel-Frequency Cepstral Coefficients

FFT is the optimal way to calculate the Discrete Fourier Transform (DFT: Bracewell & Bracewell, 1986), where DFT accounts for the frequency of a discrete signal in the frequency domain format. Basically, it converts discrete signals into the corresponding frequency domain (Oppenheim, Schafer, & Buck, 1999).

However, since there are several implementations of FFT that are already available, the representations relative spectra (RASTA) speech processing package (Hermansky, 1990; Hermansky & Morgan, 1994), was used in this study. The author proposed several audio feature extraction algorithms such as: MFCC and perceptual linear predictive (PLP) Cepstra, and the implementations are established based upon Malcolm Shaney's Auditory Toolkit (Slaney, 1998), one of the most well-known auditory model implementations. In the next chapter, the way in which the RASTA package was applied to extract audio features will be explained.

4.3.5 Audio Feature Discretization

The audio feature should be discretized in order to be compatible with the HMM, as the HMM is a discrete model. As a result of the discrete audio feature extracted, a histogram with 400 columns was utilized, and then corresponding column numbers were replaced by actual values.

4.4 Hidden Markov Model

The HMM is a probabilistic model presenting a sequence of observations (Ghahramani, 2001). Time is an important factor in HMM, as the observations are meaningful during that time. However, time may be demonstrated in many different ways, and most studies use it in equal time intervals. These observations are discrete values describing the prediction results. As time passes, the system transitions from one state to another and these transitions between visited *states* generate a *sequence of states*. According to the Markov Chain property, the last state of a subsequence is the only factor affecting the probability of the next state. In other words, the probability of a sequence is described by the conditional probability of the current and the next states.

4.4.1 Hidden Markov Model through Example

The following section describes the basic concepts of the HMM with an example in order to help the reader to better understand HMM. However, since this example does not illustrate the entire aspects of the model, it can be helpful to describe on a small scale. In this study, the audio feature vectors are observations and the markers are the hidden states. Due to the huge size of HMMs in this study, this example is introduced to illustrate the mechanism of HMM.

There is a prisoner who wishes to know what the weather is outside, but his prison cell does not have a window. While he is unable to observe the weather changes directly, his prison cell is located in front of a vending machine, and the correctional officer visits the vending machine every day and selects a cup of hot café or a bottle of orange juice or a bar of chocolate. The prisoner observes the correctional officer’s choices each day and surmises that there is a direct correlation between his choices and the daily weather. Therefore, the observations are the vending choices and the hidden states are the actual weather conditions (*cold* or *warm*).The prisoner has created a HMM, displayed below in Figure 4.4:

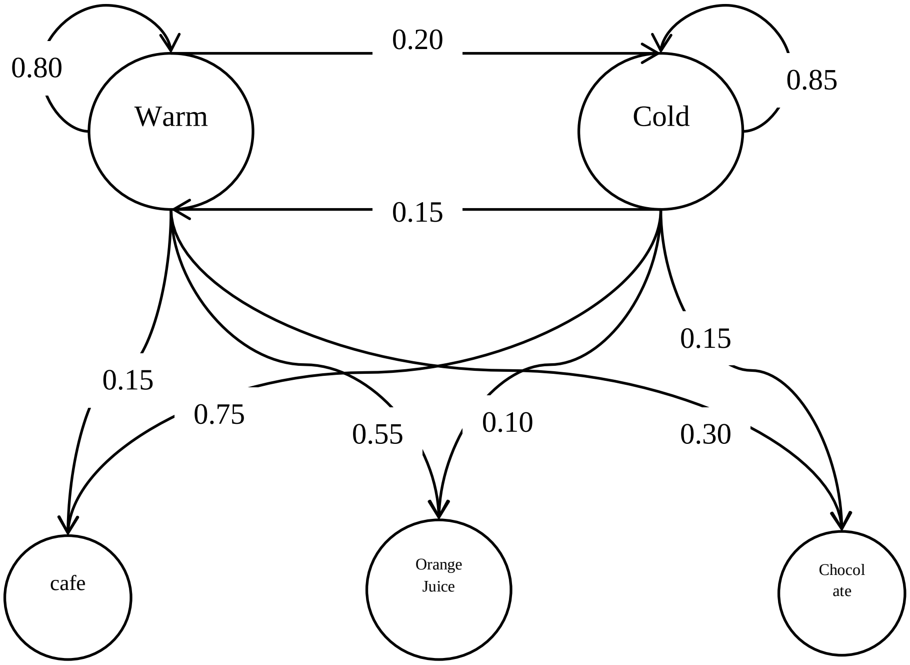


Figure 4.4 Weather Prediction Model

According to Figure 4.4, the prisoner may surmise what the weather may be based upon the observation from the vending machine. There are two types of transition in this diagram that describe the transition between hidden states observations and internal transition between hidden states. A set of states is defined by {S_1, S_2, …, S_n} and in this example the set of hidden states is {*warm*, *cold*}. The probability of each sequence depends upon the condition probability of its current and the previous states showing in Eq.4.5.

$$P(S_k|\, S_1, S_2, S_3, \dots ,S_{k-1}) = P(S_k|\, S_{k-1}) \qquad (4.5)$$

The starting point of the calculation is still unclear, as the probability of starting from state *i* and state *j* are different. Consequently, an initial state is needed to determine the probability of the next visiting state. In this example, the prisoner connects a point at the top level as an initial state $\pi = P(S_i)$, as is shown in Figure 4.5.

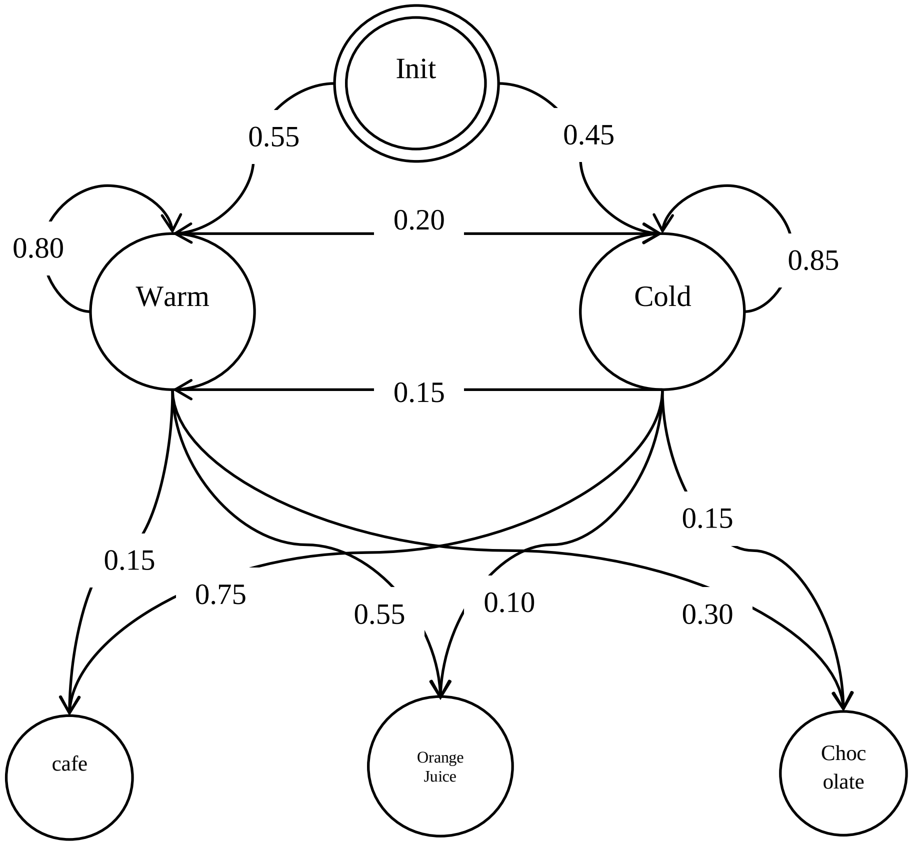


Figure 4.5 Weather Prediction Model with an Initial State

According to the Figure 4.5, the state *transition probability matrix* would be as follows:

$$\begin{pmatrix} 0.80 & 0.20 \\ 0.15 & 0.85 \end{pmatrix}$$

The transition probability for cold weather if yesterday was cold is shown by Eq.4.6:

$$\mathrm{P}(Cold \mid Cold) = 0.85 \tag{4.6}$$

The *initial state* would be similar to a matrix showing the transition from initial state to the hidden states.

$$\begin{pmatrix} 0.55 \\ 0.45 \end{pmatrix}$$

According to the Markov chain property, the generic format is found by Eq.4.7:

$$P(S_1, S_2, S_3, \dots ,S_k) = P(S_k| S_1, S_2, S_3, \dots ,S_{k-1})\ P(S_1, S_2, S_3, \dots ,S_{k-1}) = \dots$$

$$= P(S_k|S_{k-1})\ P(S_{k-1}|S_{k-2})\ P(S_{k-2}|S_{k-3}) \dots P(S_2|S_1)\ P(S_1) \tag{4.7}$$

This formula estimates the probability of a sequence of hidden states; yet, the observations have not been considered, and another matrix may determine the transition probability from hidden state to visible states called the *emission probability matrix*.

$$\begin{pmatrix} 0.15 & 0.55 & 0.30 \\ 0.75 & 0.10 & 0.15 \end{pmatrix}$$

The generated observation matrix above displays the transition between hidden and observation states. In which the above matrix, the sum of horizontal probability values in each line is equal to one.

The prisoner example did not involve the concept of training, because it was assumed that the prisoner knew the transition probabilities for both matrices. Most problems did not provide pre-knowledge, thus the learning issue was a concern. The Forward-Backward algorithm calculates M model properties with a sequence of observations and corresponding states. In this study, discretized audio features plays the role of prisoner's observations and location of markers are hidden states, consequently; according to Figure 4.5 in this proposed model, the bottom nodes represent observation (discretized audio vectors) and middle nodes illustrate hidden states (markers).

4.4.2 Evaluation, Decoding and Learning Problems in Hidden Markov Model

Assuming the model *M* and the observation sequence O = {O_1, O_2, O_3, … , $O_{N,}$}, where $O_i \in$ observation set V={V_1, $V_{2,}$ …, V_k}, estimates the most probable sequence. Computing the probability for all hidden states with the assumed observation sequence is an exponential complexity problem. In order to overcome this issue, dynamic programming may be applied to reduce it to cubic time complexity. Hence, the best way to compute it is with the Forward-Backward algorithm (Yu & Kobayashi, 2003), though this approach may also applicable for determining the most likely sequence of hidden states through a given sequence of observations. The Viterbi algorithm (Forney Jr, 1973) is another well-known algorithm for solving the decoding problem by using the Forward-Backward algorithm. In this study, a Forward-Backward algorithm was applied to estimate the emission and transaction matrices. The Viterbi algorithm was applied for finding the most likelihood sequence of states.

4.4.2.1 Evaluation and estimating the Hidden Markov Model matrices

The Forward-Backward algorithm (Yu & Kobayashi, 2003) used was a perfect example of dynamic programing (Bellman, 1956; Howard, 1960) and is described briefly below:

Assumptions as known:

- Emission probability: P(V_k| S_k)
- Transition probability distribution: P(S_k| S_{k-1})
- Initial states: P(S_1)

Notations:

- S = (S_1, S_2, S_3, … ,S_n)
- $S_{i:j}$ = (S_i, S_{i+1}, S_{i+2}, … ,S_j)

This algorithm should calculate P(S_k| $V_{1:n}$), and is split into two parts, which consist of the Forward algorithm which calculates P(S_k| $V_{1:k}$) ∀ $K = 1,\dots,n$. and the Backward algorithm which computes P($V_{k+1:n}$| S_k) ∀ $K = 1,\dots,n$ and consequently Eq.4.8:

$$P(S_k | V_{1:n}) \propto P(S_k, V_{1:n}) = P(V_{k+1:n} | S_k , V_{1:k})\, P(V_k | S_{1:k}) \qquad (4.8)$$

Given S_k, $V_{1:k}$ and $V_{k+1:n}$ are conditionally independent, therefore Eq.4.9:

$$P(S_k | V_{1:n}) = P(V_{k+1:n} | S_k)\, P(V_k | S_{1:k}) \qquad (4.9)$$

As is shown in the above formula, P(S_k| $V_{1:n}$) is the multiplication of Forward and Backward algorithms.

4.4.2.2 Forward Algorithm

As shown in Eq.3.8, the purpose of the Forward algorithm is computing P(S_k , $V_{1:k}$) in Eq.4.10.

$$P(S_k , V_{1:k}) = \sum_{z_{k-1}}^{m} P(S_k, S_{k-1}, V_{1:k}) =$$

$$\sum_{z_{k-1}}^{m} P(V_k|S_k, S_{k-1}, V_{1:k-1}) P(S_k|S_{k-1}, V_{1:k-1}) P(S_{k-1}, V_{1:k-1}) \qquad (4.10)$$

Due to S_k and S_{k-1} are conditionally independent of S_{k-1} , V_k and V_{k-1}, thus Eq.4.11:

$$P(S_k , V_{1:k}) = \sum_{z_{k-1}}^{m} P(V_k|S_k) P(S_k|S_{k-1}) P(S_{k-1}, V_{1:k-1}) \qquad (4.11)$$

Here as assumed, the first part is the emission probability, the second part is Transition Probability, and the last part is $\alpha_k(S_k)$. So the above formula is re-defined by Eq.4.12:

$$P(S_k , V_{1:k}) = \sum_{z_{k-1}}^{m} P(V_k|S_k) P(S_k|S_{k-1}) \alpha_{k-1}(S_{k-1}) \qquad (4.12)$$

As called P(S_k , $V_{1:k}$). = $\alpha_k(S_k)$ and then the recursion would be Eq.3.13 and Eq.4.14:

$$\alpha_k(S_k) = \sum_{z_{k-1}}^{m} P(V_k|S_k) P(S_k|S_{k-1}) \alpha_{k-1}(S_{k-1}) \qquad (3.13)$$

$$\alpha_1(S_1) = P(S_1, V_1) = P(V_1|S_1) \qquad (3.14)$$

4.4.2.3 Backward Algorithm

This algorithm is exactly the same as the Forward algorithm, just in the opposite direction where Eq.3.15 and Eq.3.16 illustrate computing P($V_{k+1:n}$ |S_k) for all of the rest of the values from k to the end:

$$P(V_{k+1:n}|S_k) = \sum_{z_{k+1}}^{m} P(V_{k+1:n}, S_{k+1}|S_k) \tag{4.15}$$

$$P(V_{k+1:n}|S_k) = \sum_{z_{k+1}}^{m} P(V_{k+2:n}|S_{k+1}, S_k, V_{k+1}) P(V_{k+1}|S_{k+1}, S_k)\, P(S_{k+1}|S_k) \tag{4.16}$$

Due to V_{k+2} is conditionally independent of S_k , V_{k+1} given S_{k+1} thus Eq.4.17 would be:

$$P(V_{k+1:n}|S_k) = \sum_{z_{k+1}}^{m} P(V_{k+2:n}|S_{k+1}) P(V_{k+1}|S_{k+1}, S_k)\, P(S_{k+1}|S_k) \tag{4.17}$$

In order for V_{k+1} is conditionally independent of S_k given S_{k+1} thus in Eq.4.18:

$$P(V_{k+1:n}|S_k)= \sum_{z_{k+1}}^{m} P(V_{k+2:n}|S_{k+1}) P(V_{k+1}|S_{k+1})\, P(S_{k+1}|S_k) \tag{4.18}$$

Then $P(S_{k+1}|S_k)$ is called β_k (V_k) defining in Eq.4.19. The same as the above:

$$\beta_k(S_k) = \sum_{z_{k+1}}^{m} P(V_{k+1}|S_{k+1}) P(S_{k+1}|S_k) \beta_{k+1}(S_k) \tag{4.19}$$

4.4.3 Decoding of the Model

One of the most well-known algorithms used for HMM decoding has been the Viterbi algorithm (Forney Jr, 1973; Hagenauer & Hoeher, 1989). The Forward-Backward algorithm has been applicable in decoding so that the basic idea of this algorithm has been as a recursive process for finding the most probable sequence of hidden states where the sigma in Forward recursion formula was removed and the highest probability was determined to be the answer.

To compute the most likely sequence $\delta = argMax\ P\ (S|V)$ should be calculated where $\delta = argMax\ P\ (S|V)$ and S = (S_1, S_2, S_3, … ,S_n) and V = (V_1, V_2, V_3, … , V_n).

$\max_{x,y} f(x)g(x,y) = \max_x [f(x) \max_y g(x,y)]$ if $f(x) \geq 0\ for\ all\ x\ values\ and\ g(x,y) \geq 0\ far\ all\ x\ and\ y\ values.$ And $\max_S P(S|V) = \max_S P(S,V)$.

Based upon the Markov Chain property Eq.4.20 is written:

$$\mu_k(S_k) = \max_{S_{1:k-1}} P(S_{1:k}, V_{1:k}) = \max_{S_{1:k-1}} P(V_k|S_k)\, P(S_k|S_{k-1}) P(S_{1:k-1}, V_{1:k-1}) \qquad (4.20)$$

Then $f(x)$ is used in Eq.4.21:

$$\mu_k(S_k) = \max_{S_{1:k-1}} [\, P(V_k|S_k)\, P(S_k|S_{k-1}) \max_{S_{1:k-2}} P(S_{1:k-1}, V_{1:k-1})] \qquad (4.21)$$

So recursively, the max may be calculated and determined, as the equation is the multiplication of emission probability, transition probability, and recursion part ($\mu_{k-1}(S_{k-1})$).

4.4.4 Learning Issue and Estimation of Hidden Markov Model Parameters

In this study, Maximum Likelihood (Gales, 1998; Leggetter & Woodland, 1995), was applied to estimate the model parameters, mainly consisting of transition and emission matrices. Consequently, maximizing the probability of a given arbitrary sequence would be the purpose. No analytical solution was introduced to solve the maximum likelihood problem for the whole model, but there were several algorithms which were applied to find one of the local maximums by iterative calculations. Obviously, there was no guarantee that a designated local maximum was the explicit maximum.

CHAPTER 5: THE MODEL

5.1 Magnetic Resonance Imaging[3]

A fast-gradient echo FLASH (Fast Low Angle Shot) multi-shot spiral technique was used to acquire 15.8 frames per second (fps) of the midsagittal image plane during the production of ”ansa.” The speech sample was chosen to represent movements of the velum between fully lowered (i.e., nasal), elevated (i.e., consonants), and transitions between both positions. A metronome beat of 2 Hz was played over head phones to control the rate of the speech tasks (one syllable per beat). This imaging speed allowed for at least one full image during each lowered and each elevated production to analyze the data for a nasal and oral sound.

The imaging sequence used a time-efficient acquisition of a six-shot spiral pulse sequence with an alternating TE between 1.3 and 1.8 ms to allow for dynamic estimation and correction of the magnetic field map. Saturation bands were used to suppress the signal from regions outside of the area of interest and to provide greater separation between higher fat concentrations areas such as the cheeks. Fast frame rates were achieved through the use of an optimized acquisition strategy coupled with an image reconstruction method that corrected for effects caused by imperfections in the magnetic field in the oropharyngeal region (Sutton, Conway, Bae, Seethamraju, & Kuehn, 2010).

[3] Images were provided by Department of Communication Sciences & Disorders at East Carolina University. Data acquisition and the corresponding information were reported in (Perry, Kuehn, Sutton, & Gamage, in press). The author and Computer Science Department at East Carolina University did not participate in data acquisition process.

Images were reconstructed with an output time-driven sliding window process at 40 frames per second (fps). This process allowed data to have a minimal amount of interpolation across time and uses the native frame rate (15.8 fps) to interpolate images to the desired output rate. The sliding window reconstruction process minimized redundant information in adjacent time points and minimizes temporal blurring (Sutton et al., 2009).

Acquisition simulation software provided by the vendor of the MRI scanner provided timing data which was used to align the audio speech recordings with the dynamic images. This software allowed for accurate simulations of sequence timing using the exact acquisition protocol, providing information about the actual time location of data acquisition events with 10s accuracy.

5.2 Feature Extraction and Audio Signal

In chapter 3, the audio feature extraction was explained. The audio features were extracted from the stream of audio signal after conducting noise cancelation filters. Despite the speech recognizers, the audio signal was not segmented by VAD. Consequently, the audio was synchronized by visual features extracted. Figure 5.1 demonstrates the training process of the proposed model including two paths for audio and visual feature extraction. Both paths were used for the training purpose, provided that Figure 5.2 accounts for the prediction process based on audio signal analysis.

MRIs were imported into a visual and motion graphic software program[4] where the images sequence of entire 45 seconds was cropped to select a region of 7.5 seconds sequence.

[4] Adobe After Effects, CS 6, Adobe Systems

The audio and the image sequences were isolated from 7.5 video at 16 KHz and 40 frames per second respectively in order to produce a number of frames per second to be a factor of 1:400 sampling rate. The audio feature extraction was accomplished by using MFCC.

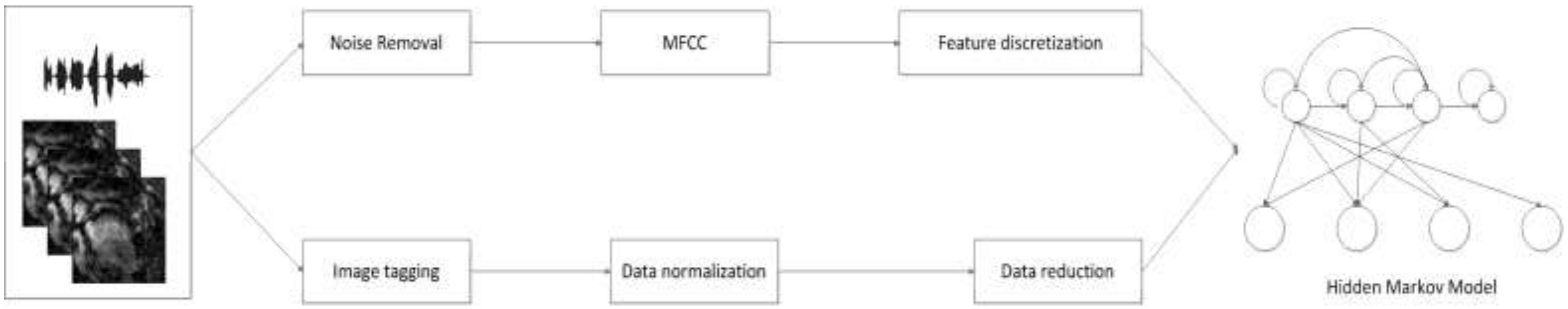


Figure 5.1 Training Phase of Proposed Model considering Audio and Visual Feature Extraction Paths

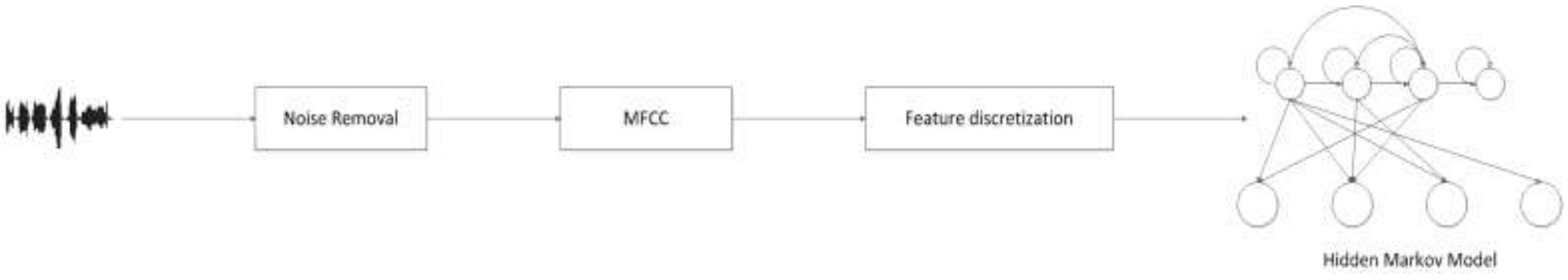


Figure 5.2 Prediction Phase of Proposed Model

The MFCCs are short term spectral-based features (Li et al., 2000) where the Mel-scale is chosen close to the human auditory system (Slaney, 1998). In this study, the final acoustic feature dimension was 39, including MFCC coefficient transformations (with 13 elements) and the first and the second derivatives (13 elements for each derivative). The extracted audio features were discretized and labeled in 400 distinct classes from 1 to 400.

5.3 Visual Feature Extraction by Proposed Tagging Method

Visual features were extracted using the MR images sequence by selecting four markers along the nasal surface of the velum and three markers along the posterior pharyngeal wall. One stationary pivot point was placed at the posterior nasal spine (PNS). As shown in Figure 5.3, the markers were positioned so that the third marker was located at the velar knee, and fewer markers were used along the uvula as a result of the lack of significance in this region during speech production. Figure 5.4 demonstrates corresponding labels of tags over the velum and pharyngeal wall.

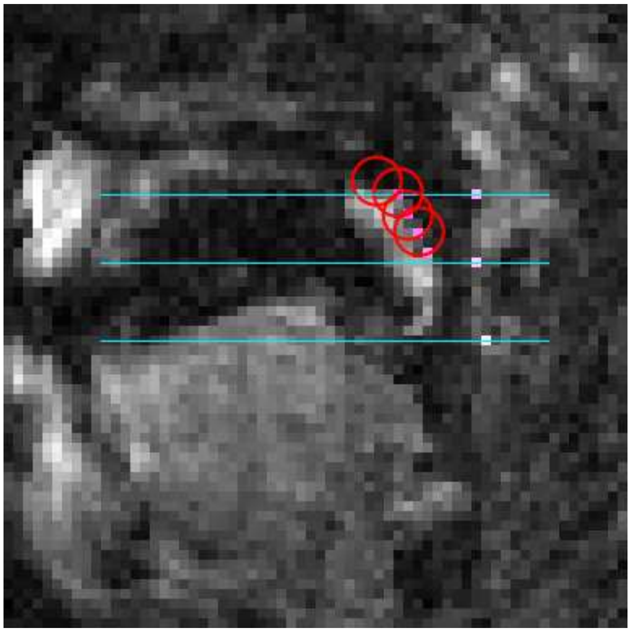


Figure 5.3 Selected Markers along the Nasal Surface and Posterior Pharyngeal Wall

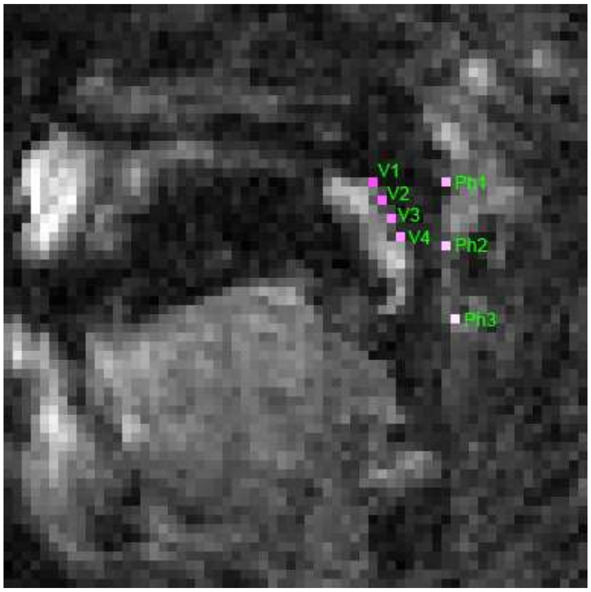


Figure 5.4 Velar Markers Demonstrated by Letter *V* and Pharyngeal Wall Labeled by Letters *Ph*

In order to keep a consistent distance between each marker, a circular tracking tool was created to identify markers along the length of the velum using the initial marker placed at PNS. The program was designed to then draw a circle with a 13 pixels radius around the initial stationary pivot-point. The second marker was then marked at the intersection of the circle with the nasal velar surface. This method, as shown in Figure 5.5, demonstrated the repetition of this process to identify every positioned marker along the velar surface.

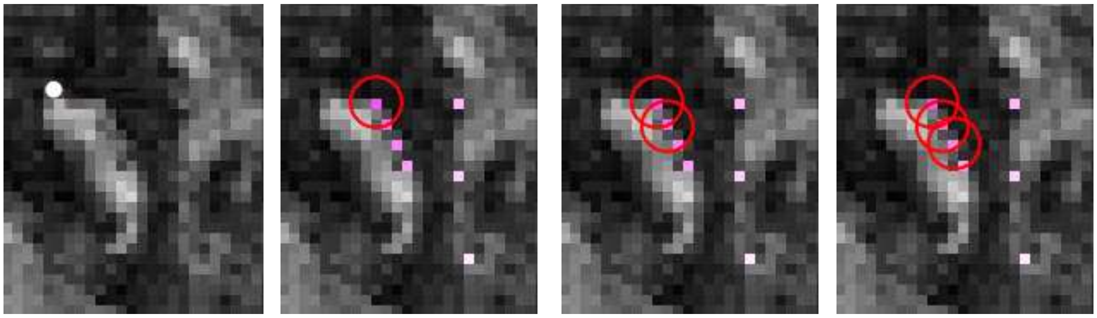


Figure 5.5 Makers at the Intersection of the Circle with Nasal Velar Surface

Anterior and posterior pharyngeal wall movements were calculated in the horizontal (x-axis) dimension (Figure 5.6). The first horizontal line was placed 24 pixels below of PNS. The second and third lines were placed 48 and 96 pixels below the first line respectively. Horizontal lines were created to marker out pharyngeal markers with the same distance. The closest on the intersection of pharyngeal wall and each line was tagged by the researcher.

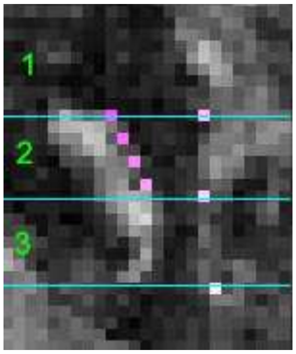


Figure 5.6 Pharyngeal Wall Markers

Three hundred sequential images were tagged by the researcher and the result was a table that consisted of 300 tuple (rows), including 7 markers (4 velar and 3 pharyngeal), and 14 columns. Each marker demonstrated movement in both the x and y-axis, yielding two values for each marker. For each marker, the x-value was multiplied by 1000 and then the corresponding y-value was added to create the concatenation of the x and y columns. This new set of numbers was labeled from 1 to n which represented the total distinct classes for the compounded location of the x and y-axis for the marker across the speech sample. In order to reduce the number of classes and simplify the model (e.g. HMM), each class of the set equaling less than 10% of total samples, was merged with the next class iteratively.

5.4 Creating Hidden Markov Models

The HMM was used (L. Rabiner, 1989) to predict the velar and pharyngeal wall boundaries. The audio feature extracted was the observation (inputs) of HMM and visual features were considered to be the internal hidden states (outputs) of the HMM. In contrast to linear left to right HMM in speech recognition systems (Ghitza, 1994), the topology of the model developed in this study did not follow a linear pattern. Two parameters of the model consisted of transition and emission matrices (L. Rabiner, 1989) that were estimated based upon the visual and audio features. The transition probability matrix found the path from a hidden state at time *t* by providing the hidden state at time *t-1*. For each possible element of the class (N elements), there were N possibilities to transition from *t-1* to *t*. Thus, the N×N matrix was estimated for each marker by using the corresponding number of possibilities (hidden states).

The number of audio features (observations) was the same for all HMM trainings. The emission matrix was N×400 for each marker (Table 5.1). The HMM parameters were estimated for each marker using Matlab[5] software. To prevent the transition probability from being zero, all zero-elements in emission and transition matrices were replaced by a small number (e $=10^{-7}$). The models were trained using a 200 audio feature data set and 100 samples were set aside for testing. The Viterbi algorithm (Forney, 1973), was then applied to predict the most likely sequence of hidden states.

	V1	V2	V3	V4	Ph1	Ph2	Ph3
Hidden States	4	8	18	15	3	2	2

Table 5.1 Number of Hidden States for Markers

[5] http://www.mathworks.com/products/matlab/

CHAPTER 6: RESULTS

The location and shape of the velum and pharyngeal wall were predicted at 2.5 seconds for one set of images and the results were analyzed using two distinct methods, namely *accumulative minimum distance* and *evaluation by inspection*. The former evaluation method was a mathematical approach to accumulate minimum distances between prediction and the actual corresponding markers tagged by the researcher. Although the proposed tagging method in this paper avoided arbitrary tagging by the researcher, it was not sufficient to prevent researcher introduced errors in tagged images. In other words, for one specific image, more than one set of markers has been introduced. Although the alternative set of markers may be acceptable, the current method does not support multi-markers, requiring the researcher to choose just one of the alternative markers. Consequently, the *accumulative minimum distance* may not represent an accurate measurement, where the actual acceptable prediction may be at least equal or greater than *accumulative minimum distance*. To overcome the inaccuracy issue, one pixel threshold was assumed, because the prediction residual of less than one pixel was not visible in the visualization phase.

6.1 Accumulative Minimum Distance

Figures 6.1 - 6.4 demonstrate the residual of the velar prediction result. Each point in the graphs accounts for the average error per pixel of five consecutive marker predictions. Assuming the threshold for these graphs, no error was introduced in V1 prediction, while V2, V3, and V4 were predicted with a different level of error. The error in V1 was dramatically less than other markers located on the velar surface, because four hidden states were defined for V1 (distinct classes in Table 5.1), provided that V2, V3, and V4 were defined by 8, 18, and 15 hidden states respectively. Hence, having fewer hidden states contributed to less residual in the prediction result.

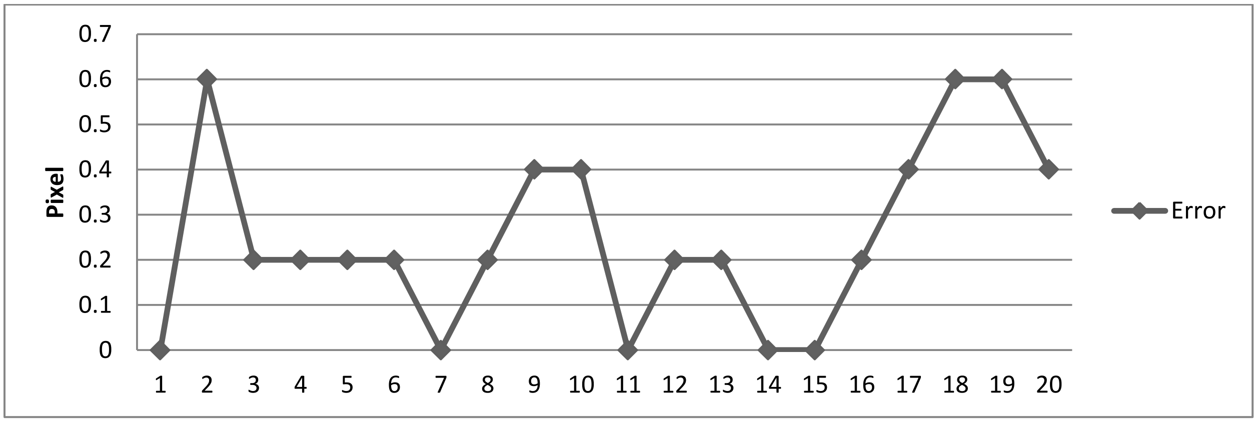


Figure 6.1 Average Errors for each Five Consecutive Predictions for *V1* Marker

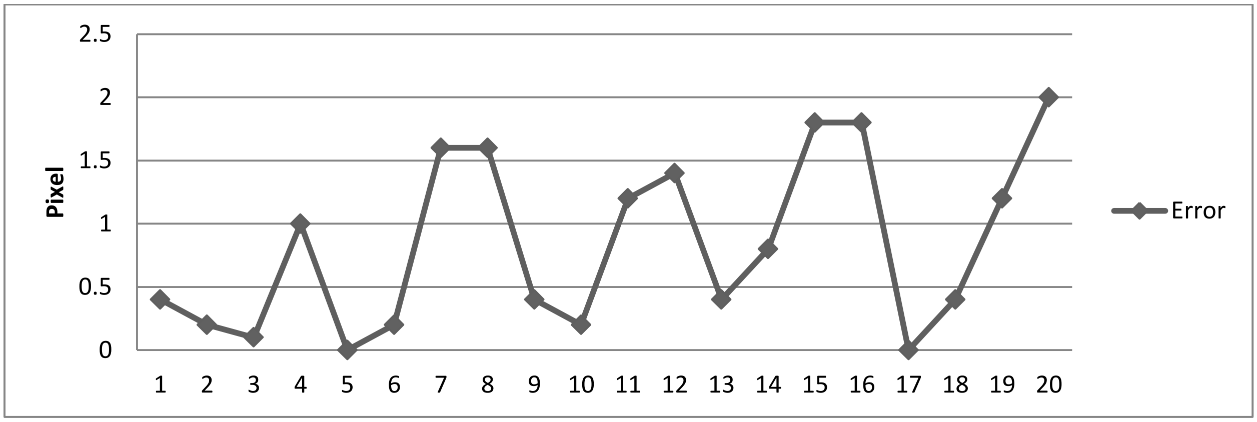


Figure 6.2 Average Errors for each Five Consecutive Predictions for *V2* Marker

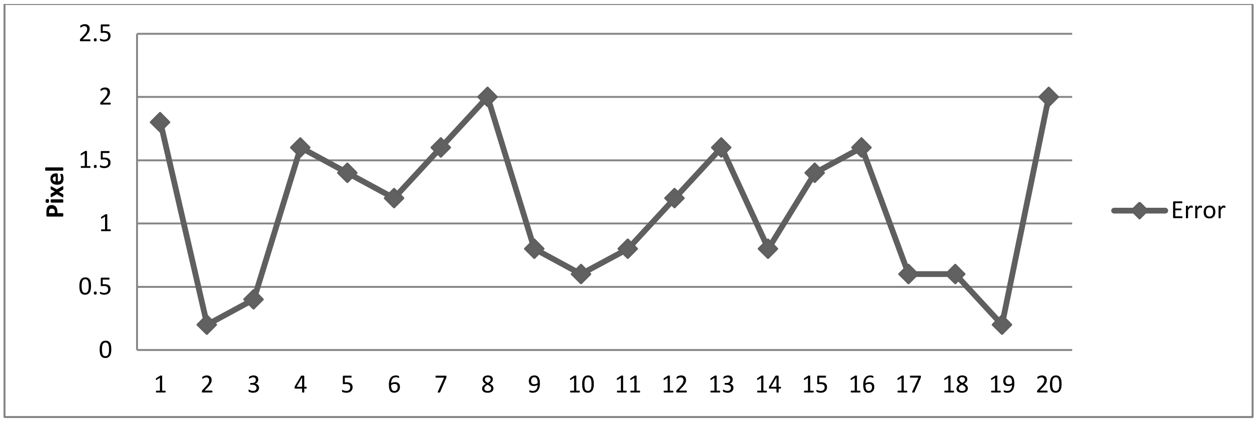


Figure 6.3 Average Errors for each Five Consecutive Predictions for *V3* Marker

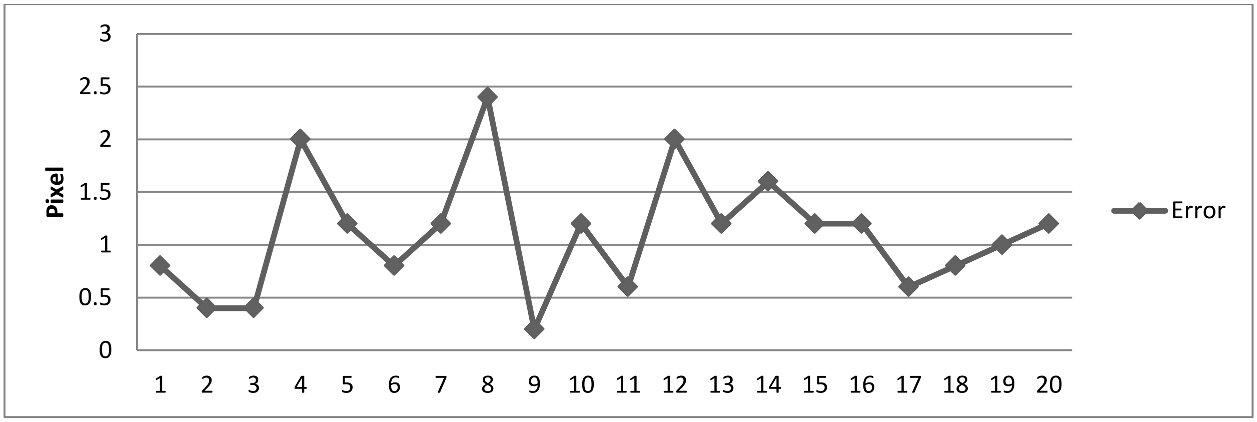


Figure 6.4 Average Errors for each Five Consecutive Predictions for *V4* Marker

Due to minor movement of the pharyngeal wall, the residual prediction introduced a maximum of one pixel error; consequently, there was no error displayed in pharyngeal wall prediction when assuming the one pixel threshold.

Figure 6.5 is an accumulative graph of error values for four markers located on the velum, where the vertical axis demonstrates the sum of error per pixels through the velum markers, and the horizontal axis accounts for the error rate average for the last 10 consecutive marker errors. However, since one pixel error was assumed to be the threshold for each marker, in Figure 6.5, the threshold was not applied because of better visibility.

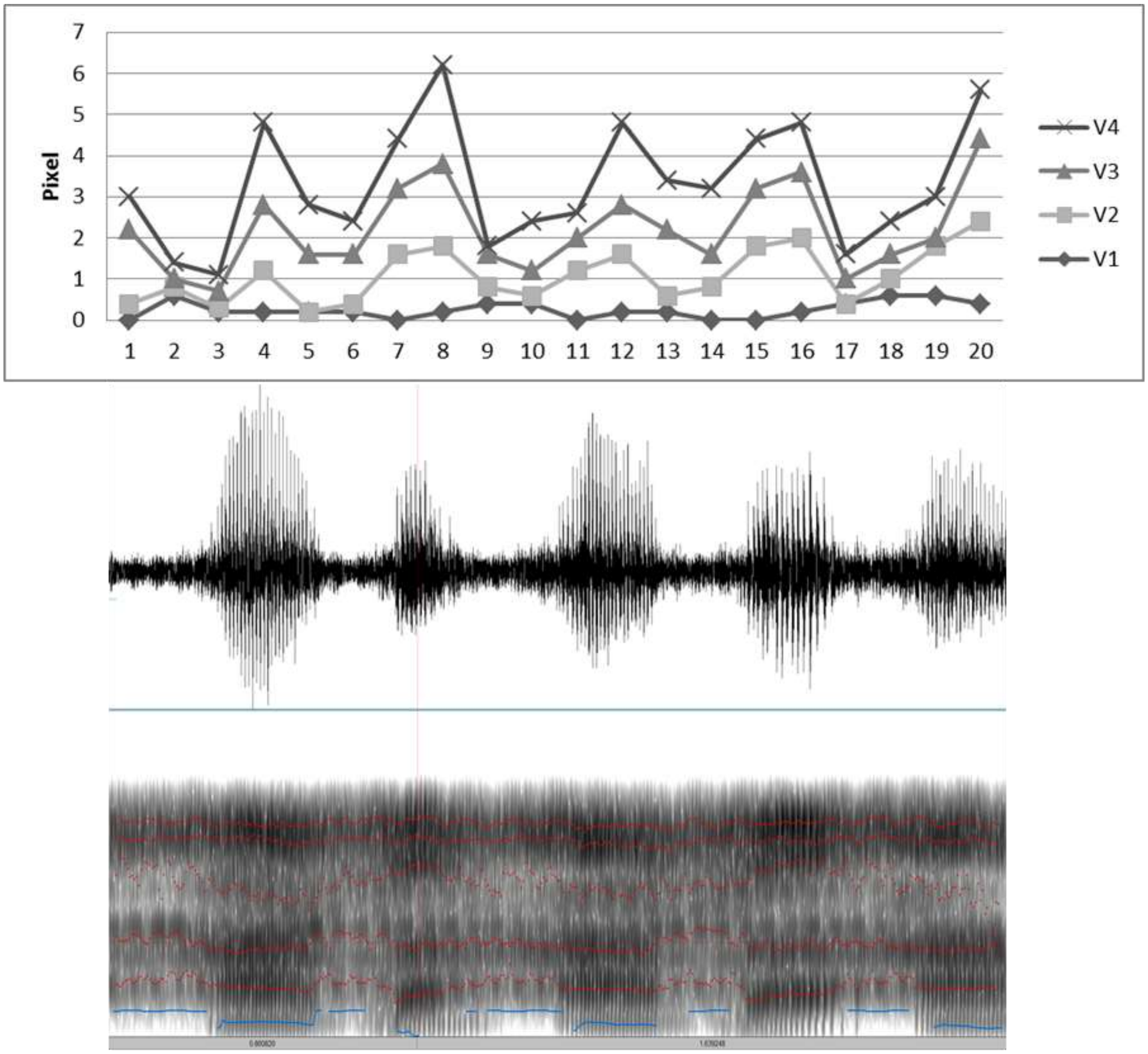


Figure 6.5 Velum Predictions Accumulative Error and Spectrogram with Formants

The visible error in the visualization phase actually would be those values by more than four pixels in the graph. This evaluation presents 81% accuracy considering one pixel threshold for 4 markers on velum.

Conceivably, there are two different interpretations regarding error variation:

- A correlation between audio signal amplitude and error rate:

  As shown in Figure 6.5, there is a correlation between accumulative error and audio signal amplitude where the higher the amplitude generated, the higher the residual introduced.

- Higher velocity of velum causes a high error rate during speech production:

  In order to produce high amplitude signals, the velum must move very fast and it has to have contact with the pharyngeal wall in a very short duration. This velocity and high level of deformability within such a short duration may be the cause of the growth in error rate. Thus, the second interpretation introduces two concerns regarding the issue, including a fewer number of samples during the velum closure, and a variety of movements and figures in a very short duration.

  The distribution of a training set for all classes did not follow the uniform distribution pattern, and therefore there is a variety of numbers of training sets for classes while the proposed model had the same policy for any input stream. In other words, classes in low amplitude domain had a higher number of training samples and they were trained well provided that high amplitude samples suffered from a low number of samples. Unbalanced training sets may be one of the reasons, but the error rate also was exacerbated by the high level of deformability of the velum during high amplitude voice production. This phenomenon dramatically fragmented these small numbers of samples to different classes. Consequently, these classes were not trained well. The performance of the HMM has been improved in many studies with a variety of signals and noises, so the fewer numbers of sampling and normal distribution of training samples (exponential shaped) would be the cause of the high error rate in some parts of the prediction.

6.2 Evaluation by Inspection

This method was evaluated by comparing the location of superimposed predicted structure figures over the velum and pharyngeal wall. The compared researcher markers and predicted markers were determined where the result was either *pass* or *fail*. However, in the superimposed set of prediction images, where research markers were included and the predicted markers were not located on those markers (Table 6.1), the whole predicted figures were acceptable because the several combinations of markers were conceivable for a single image. The result of the inspection was an 83% acceptance rate.

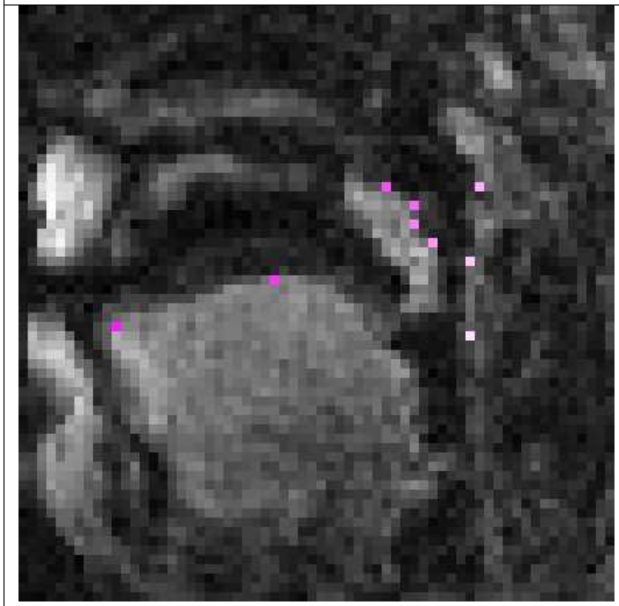
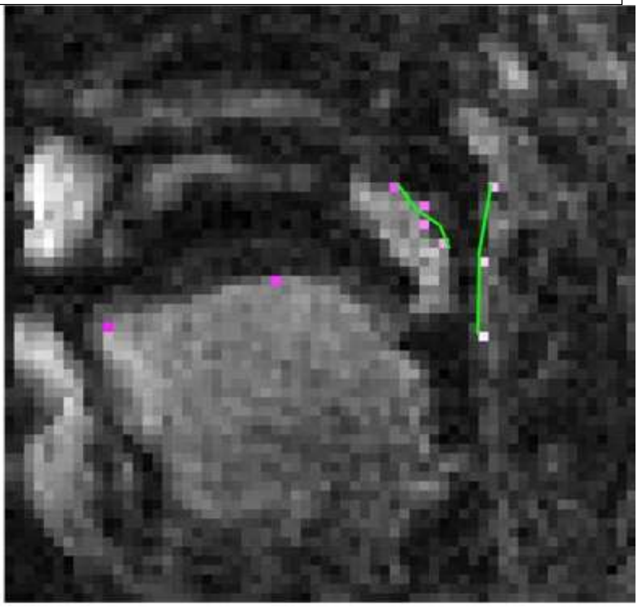
Marked by Researcher	Superimposed predictions
	

Table 6.1 Manual Markers vs. Superimposed Predicted Markers

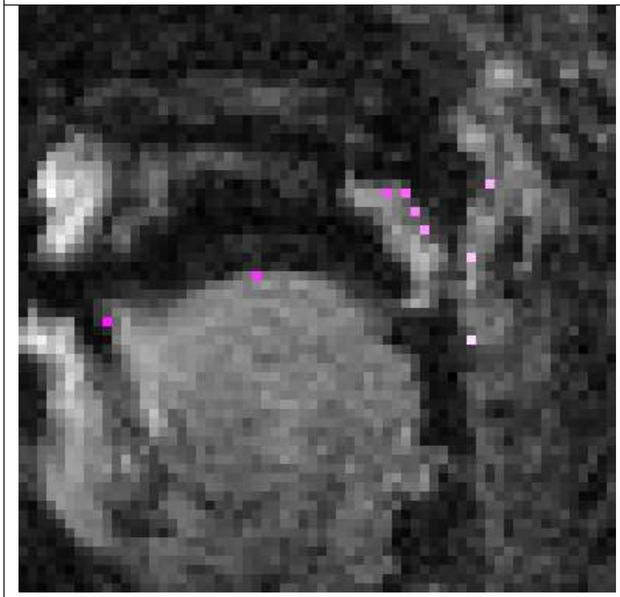
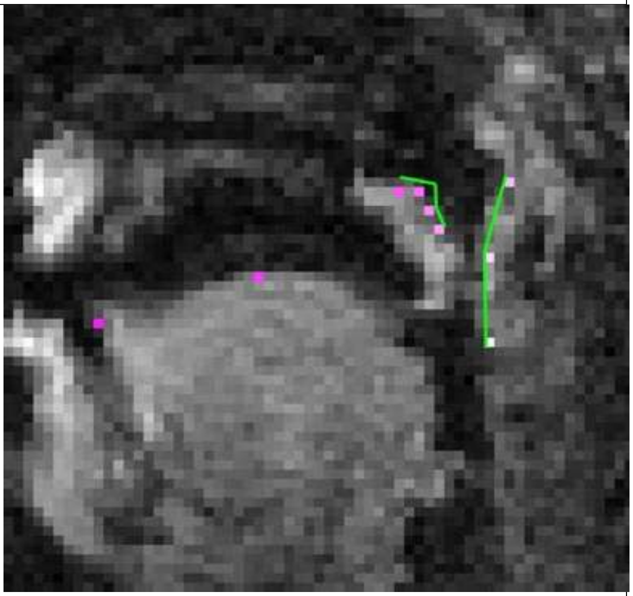
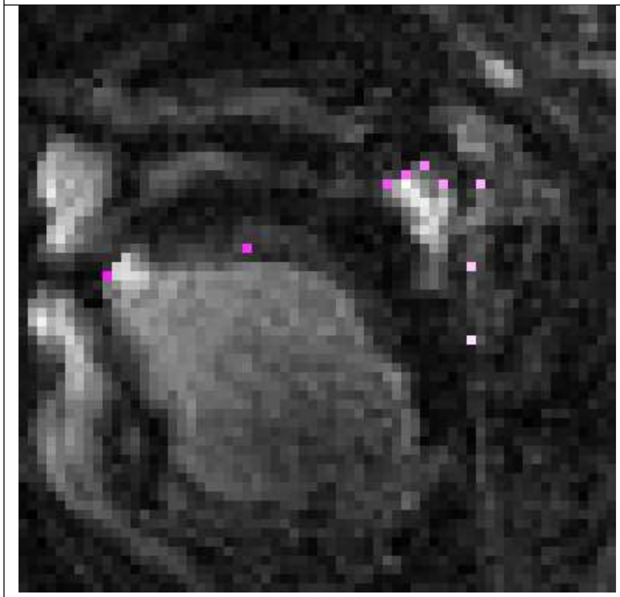
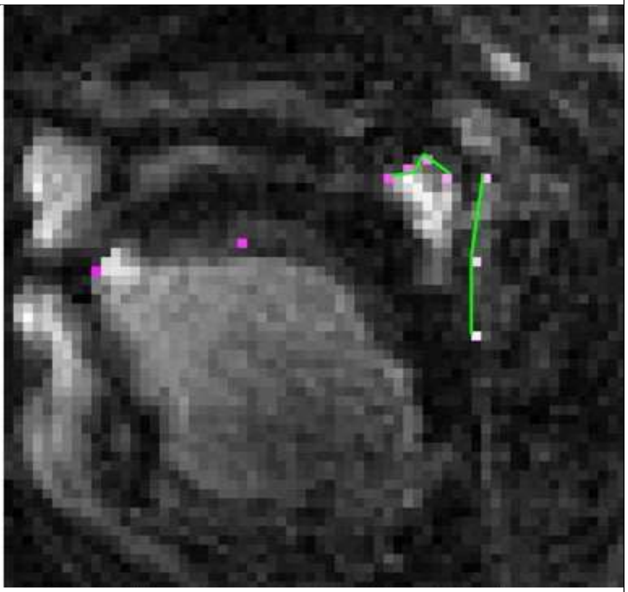
Marked by Researcher (continue)	Superimposed predictions (continue)
	
	

Table 6.1 Manual Markers vs. Superimposed Predicted Markers (Continue)

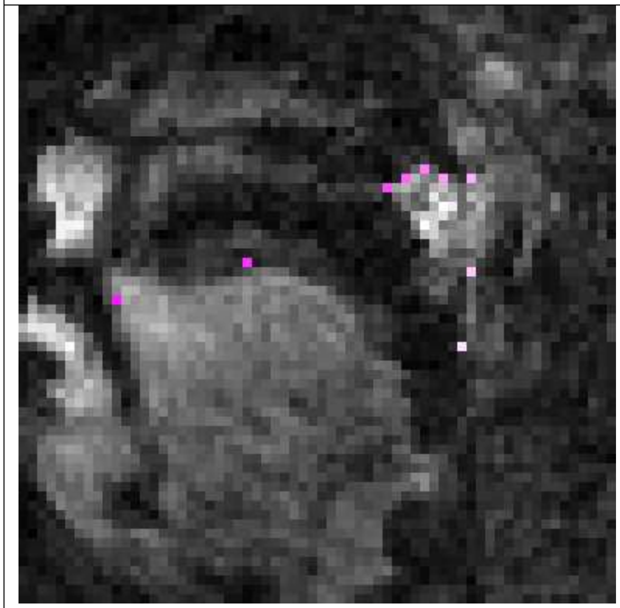
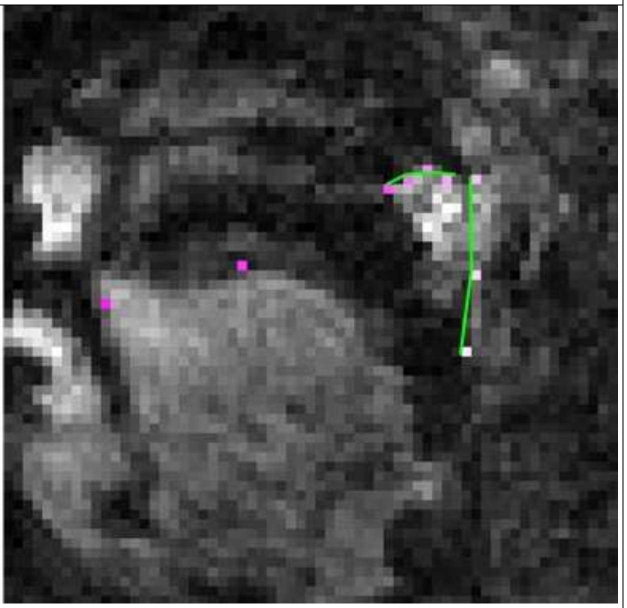
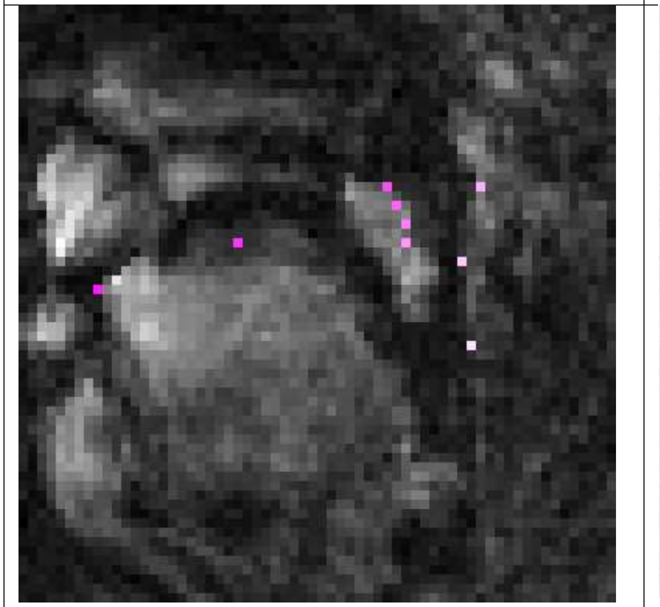
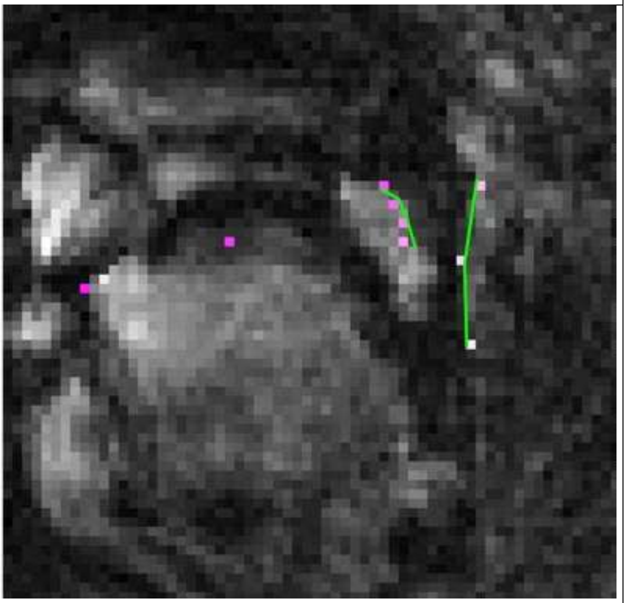
Marked by Researcher	Superimposed predictions
	
	

Table 6.1 Manual Markers vs. Superimposed Predicted Markers (Continue)

CHAPTER 7: CONCLUSION AND FUTURE WORKS

This chapter discusses the effect of the image tagging policy in terms of performance improvement of the model and the use of seven classifiers to improve the flexibility and reusability of the system. The last section of this chapter discovers a potential reason for the introduction of error and proposes a potential solution for future works.

7.1 Image Tagging and Search Space Reduction

Velum and pharyngeal wall trajectories were captured and quantified by using markers in the sequence of MR image. These markers were used for training the predictive model and the accuracy of the model was further evaluated by comparing the distances between the actual and the predicted markers. Arbitrary tagging of markers on the velum and pharyngeal wall may have contributed to a deterioration of the prediction results because of the increase in search space. Moreover, the system was expected to introduce a one-to-one map to demonstrate the correlation between a chunk of audio signal and a set of markers. In order to reduce the search space, a circular tagging method developed, was implemented to maintain an equal distance between consecutive markers along the velum boundary. This method decreased search space dimensions significantly and it improved the marker probability densities in the circumference.

To reduce observation variation, each instance (value) of X and Y having equal or less than 10 percent of all records were merged with the adjacent group. This process was performed iteratively to merge all instances in which there were less than or equal to 10 percent of all of the records. Thus, the model was limited to 100 observations (markers) because it had at most 10 distinct instances for X and at most the same number for Y. The combined circular method and data reduction contributed to a drastic reduction in the number of observations, and as shown in Table 4.1, the model with the largest number of observations was determined to be 18. Although data reduction occasionally resulted in the loss of some useful information, in this study, data reduction was achieved by merging less frequently inserted instances (markers) resulting from human error or from the low-quality of the images.

7.2 Hidden Markov Model and Flexibility Concerns

In order to predict a high level of structure deformability, seven markers were selected to describe the shape of the structures. Each marker, in terms of movement, had different trajectory characteristics. The markers located close to the hard palate had less movement than the others on the spine of velum. Prediction of all markers in one HMM created a large number of hidden states and all parameters may not be trained very well (Geiger, Schenk, Wallhoff, & Rigoll, 2010). As a result of the reduction from the HMM complexity, each marker was modeled by separate HMMs. The final prediction was generated by a combination of individual marker predictions. The Audio signal as the sequence of observation was common for all seven models and the hidden states were designed for each marker individually based upon the corresponding combination of X and Y ranges.

Using a set of templates to describe the location and figure of structure was an alternative to be considered as hidden states. The benefit of this alternative was simplicity, as the only one HMM is able to handle these templates. The training is very simple and there are no jitters in the prediction model, because the model just finds the most similar template to the actual configuration among few choices; Yet, this model is able to work with a specific utterance. The main advantage of using several classifiers is that, they are able address different combination of markers, regardless of the signal content.

7.3 Amplitude Vulnerability and Future Works

As shown in Figure 7.1, there is a correlation between high amplitude signals and error rates. Figure 6.1 reveals the distribution of input speech signal amplitude. However, the most frequent amplitude is on the zero column and its neighbors, during low amplitude speech production, but the velum had fewer movements in comparison to high amplitude and this movement required more time (had a longer duration) because of low velocity. Consequently, there were more samples of low amplitude moments (audio and video samples) and it created a well-trained model for low amplitude signals. A uniform sample distribution may solve the problem, but as a result of choosing HMM, training should be conducted considering sequence and time. The best solution to satisfy both conditions would be chunking an audio signal to the smaller segments where there would be a fairer sample distribution. In other words, the input signal should be trimmed where low and high amplitude is balanced. Another future study conceivably would be adding a post processing phase and designing a smoother filter such as a Kalman filter (Welch & Bishop, 1995) to make the prediction more accurate and to remove jitters.

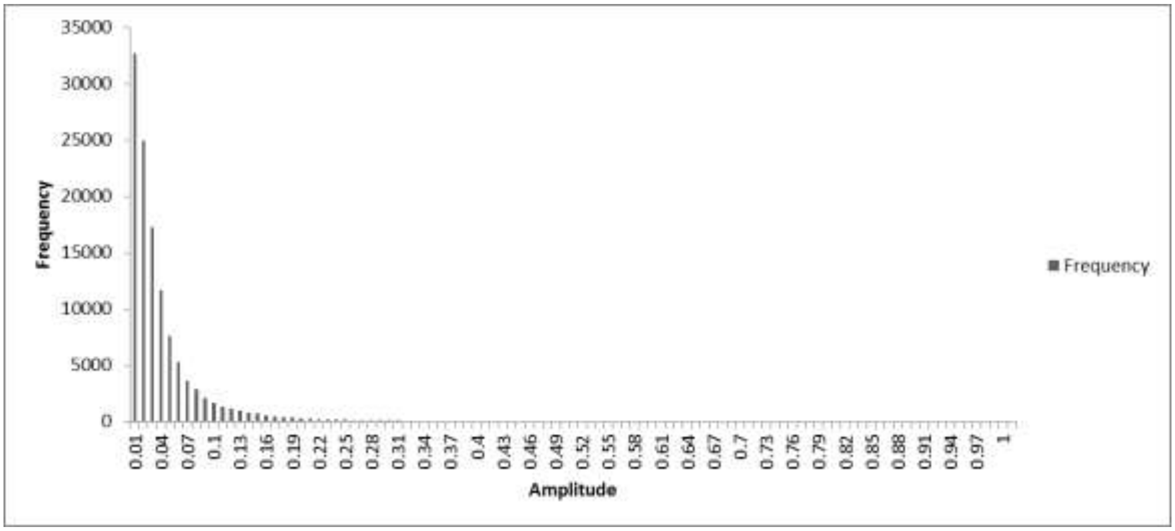


Figure 7.1 Speech Amplitude Histogram

REFERENCES

Bachu, R., Kopparthi, S., Adapa, B., & Barkana, B. *Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal*.

Bellman, R. (1956). Dynamic programming and the smoothing problem. *Management Science, 3*(1), 111-113.

Benesty, J. (2008). *Springer handbook of speech processing*: Springer Verlag.

Bracewell, R. N., & Bracewell, R. (1986). *The Fourier transform and its applications* (Vol. 31999): McGraw-Hill New York.

Braz Junior, G., Cardoso de Paiva, A., Corrêa Silva, A., & Cesar Muniz de Oliveira, A. (2009). Classification of breast tissues using Moran's index and Geary's coefficient as texture signatures and SVM. *Computers in Biology and Medicine, 39*(12), 1063-1072.

Brice, C. R., & Fennema, C. L. (1970). Scene analysis using regions. *Artificial intelligence, 1*(3), 205-226.

Campbell, R., & Dodd, B. (1980). Hearing by eye. *Quarterly Journal of Experimental Psychology, 32*(1), 85-99.

Chiappa, S., & Bengio, S. (2004). *HMM and IOHMM modeling of EEG rhythms for asynchronous BCI systems.* Paper presented at the European Symposium on Artificial Neural Networks, ESANN.

Clark, M. C., Hall, L. O., Goldgof, D. B., Velthuizen, R., Murtagh, F. R., & Silbiger, M. S. (1998). Automatic tumor segmentation using knowledge-based techniques. *Medical Imaging, IEEE Transactions on, 17*(2), 187-201.

Comaniciu, D., & Meer, P. (1997). *Robust analysis of feature spaces: color image segmentation.* Paper presented at the Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on.

Cremers, D., Rousson, M., & Deriche, R. (2007). A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *International journal of computer vision, 72*(2), 195-215.

Dalby, J., & Kewley-Port, D. (1999). Explicit pronunciation training using automatic speech recognition technology. *Calico Journal, 16*(3), 425-445.

Di Cataldo, S., Ficarra, E., Acquaviva, A., & Macii, E. (2010). Automated segmentation of tissue images for computerized IHC analysis. *Computer methods and programs in biomedicine, 100*(1), 1-15.

Due Trier, Ø., Jain, A. K., & Taxt, T. (1996). Feature extraction methods for character recognition-a survey. *Pattern recognition, 29*(4), 641-662.

Forney, G. D., Jr. (1973). The viterbi algorithm. *Proceedings of the IEEE, 61*(3), 268-278. doi: 10.1109/proc.1973.9030

Forney Jr, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE, 61*(3), 268-278.

Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech and language, 12*(2).

Geiger, J., Schenk, J., Wallhoff, F., & Rigoll, G. (2010). *Optimizing the number of states for HMM-based on-line handwritten Whiteboard recognition.* Paper presented at the Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on.

Ghahramani, Z. (2001). An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence, 15*(01), 9-42.

Ghitza, O. (1994). Auditory models and human performance in tasks related to speech coding and speech recognition. *Speech and Audio Processing, IEEE Transactions on, 2*(1), 115-132.

Greene, P. H. (1959). *An approach to computers that perceive, learn, and reason*. Paper presented at the Papers presented at the the March 3-5, 1959, western joint computer conference, San Francisco, California.

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2006). *Feature extraction: foundations and applications* (Vol. 207): Springer.

Hagan, M. T., Demuth, H. B., & Beale, M. H. (1996). *Neural network design*: Pws Pub. Boston London.

Hagenauer, J., & Hoeher, P. (1989). *A Viterbi algorithm with soft-decision outputs and its applications.* Paper presented at the Global Telecommunications Conference, 1989, and Exhibition. Communications Technology for the 1990s and Beyond. GLOBECOM'89., IEEE.

Han, W., Chan, C.-F., Choy, C.-S., & Pun, K.-P. (2006). *An efficient MFCC extraction method in speech recognition.* Paper presented at the Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America, 87*, 1738.

Hermansky, H., & Morgan, N. (1994). RASTA processing of speech. *Speech and Audio Processing, IEEE Transactions on, 2*(4), 578-589.

Holmberg, M., Gelbart, D., & Hemmert, W. (2006). Automatic speech recognition with an adaptation model motivated by auditory processing. *Audio, Speech, and Language Processing, IEEE Transactions on, 14*(1), 43-49. doi: 10.1109/tsa.2005.860349

Hong, Z.-Q. (1991). Algebraic feature extraction of image for recognition. *Pattern recognition, 24*(3), 211-219.

Howard, R. A. (1960). DYNAMIC PROGRAMMING AND MARKOV PROCESSES.

Jelinek, F. (1998). *Statistical methods for speech recognition*: MIT press.

Jianjun, Y., Hongxun, Y., & Feng, J. (2004, 18-20 Dec. 2004). *Based on HMM and SVM multilayer architecture classifier for Chinese sign language recognition with large vocabulary*. Paper presented at the Multi-Agent Security and Survivability, 2004 IEEE First Symposium on.

Jolliffe, I. T. (1986). *Principal component analysis* (Vol. 487): Springer-Verlag New York.

Junior, G. B., Paiva, A. C., Silva, A. C., & de Oliveira, A. C. M. (2009). Classification of breast tissues using Getis-Ord statistics and support vector machine. *Intelligent Decision Technologies, 3*(4), 197-205.

Kim, K.-j., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications, 19*(2), 125-132.

Kitzing, P., Maier, A., & Åhlander, V. L. (2009). Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders. *Logopedics Phonatrics Vocology, 34*(2), 91-96.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics, 7*(312), 1-26.

Kotsiantis, S., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Frontiers in Artificial Intelligence and Applications, 160*, 3.

Krishna, A., & Sreenivas, T. V. (2004). *Music instrument recognition: from isolated notes to solo phrases.* Paper presented at the Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*(1), 1-27.

Leggetter, C., & Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer speech and language, 9*(2), 171.

Levine, M. D. (1969). Feature extraction: A survey. *Proceedings of the IEEE, 57*(8), 1391-1407. doi: 10.1109/proc.1969.7277

Li, Q., Soong, F. K., & Siohan, O. (2000). *A high-performance auditory feature for robust speech recognition.* Paper presented at the Proc. ICSLP.

Liu, Z., Wang, Y., & Chen, T. (1998). Audio feature extraction and analysis for scene segmentation and classification. *Journal of VLSI signal processing systems for signal, image and video technology, 20*(1-2), 61-79.

Logan, B. (2000). *Mel frequency cepstral coefficients for music modeling.* Paper presented at the International Symposium on Music Information Retrieval.

Logan, B., & Salomon, A. (2001). *A music similarity function based on signal analysis.* Paper presented at the IEEE International Conference on Multimedia and Expo.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE, 63*(4), 561-580. doi: 10.1109/proc.1975.9792

McKinney, M. F., & Breebaart, J. (2003). *Features for audio and music classification.* Paper presented at the Proc. ISMIR.

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE, 49*(1), 8-30.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37*(1/2), 17-23.

Nattkemper, T. W., Arnrich, B., Lichte, O., Timm, W., Degenhard, A., Pointon, L., . . . Leach, M. O. (2005). Evaluation of radiological features for breast tumour classification in clinical screening with machine learning methods. *Artificial Intelligence in Medicine, 34*(2), 129-139.

Newell, A. (1981). The heuristic of George Polya and its relation to artificial intelligence.

Nilsson, M., & Ejnarsson, M. (2002). Speech Recognition using Hidden Markov Model. *Department of Telecommunications and Speech Processing, Blekinge Institute of Technology*.

Nilsson, N. J. (1969). Survey of pattern recognition. *Annals of the New York Academy of Sciences, 161*(2), 380-401.

Nunes, A. P., Silva, A. C., & Paiva, A. C. D. (2010). Detection of masses in mammographic images using geometry, Simpson's Diversity Index and SVM. *International Journal of Signal and Imaging Systems Engineering, 3*(1), 40-51.

Obermaier, B., Guger, C., Neuper, C., & Pfurtscheller, G. (2001). Hidden Markov models for online classification of single trial EEG data. *Pattern recognition letters, 22*(12), 1299-1309.

Oppenheim, A. V., Schafer, R. W., & Buck, J. R. (1999). *Discrete-time signal processing* (Vol. 5): Prentice hall Upper Saddle River.

Pearce, D. (2000). *Enabling new speech driven services for mobile devices: An overview of the ETSI standards activities for distributed speech recognition front-ends.* Paper presented at the AVIOS 2000: The Speech Applications Conference.

Perry, J. L., Kuehn, D. P., Sutton, B. P., & Gamage, J. K. (in press). Sexual Dimorphism of the Levator Veli Palatini Muscle: *An Imaging Study. The Cleft Palate-Craniofacial Journal*.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America, 24*, 175.

Qi, F.-y., & Bao, C.-c. (2006). A Method for Voiced/Unvoiced/Silence Classification of Speech with Noise Using SVM. *Acta Electronica Sinica, 34*(4), 605.

Qingzhong, L., Sung, A. H., & Mengyu, Q. (2009). Temporal Derivative-Based Spectrum and Mel-Cepstrum Audio Steganalysis. *Information Forensics and Security, IEEE Transactions on, 4*(3), 359-368. doi: 10.1109/tifs.2009.2024718

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257-286. doi: 10.1109/5.18626

Rabiner, L. R., & Schafer, R. W. (1979). *Digital processing of speech signals* (Vol. 19): IET.

Radmard, M., Hadavi, M., Ghaemmaghami, S., & Nayebi, M. (2011). *Clustering Based Voiced/Unvoiced Decision for Speech Signals.* Paper presented at the Signal Processing Symposium (SPS).

Riedhammer, K., Stemmer, G., Haderlein, T., Schuster, M., Rosanowski, F., Noth, E., & Maier, A. (2007). *Towards robust automatic evaluation of pathologic telephone speech.* Paper presented at the Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on.

Saramäki, T. (1993). Finite impulse response filter design. *Handbook for Digital Signal Processing*, 155-277.

Sharp, J. T., Goldberg, N. B., Druz, W. S., & Danon, J. (1975). Relative contributions of rib cage and abdomen to breathing in normal subjects. *Journal of Applied Physiology, 39*(4), 608-618.

Slaney, M. (1998). Auditory toolbox. *Interval Research Corporation, Tech. Rep, 10,* 1998.

Ssnderson, C., & Paliwal, K. K. (1997, 4-4 Dec. 1997). *Effect of different sampling rates and feature vector sizes on speech recognition performance.* Paper presented at the TENCON '97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications., Proceedings of IEEE.

Sutton, B. P., Conway, C., Bae, Y., Brinegar, C., Liang, Z.-P., & Kuehn, D. P. (2009). *Dynamic imaging of speech and swallowing with MRI.* Paper presented at the Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE.

Sutton, B. P., Conway, C. A., Bae, Y., Seethamraju, R., & Kuehn, D. P. (2010). Faster dynamic imaging of speech with field inhomogeneity corrected spiral fast low angle shot (FLASH) at 3 T. *Journal of Magnetic Resonance Imaging, 32*(5), 1228-1237.

Tang, C., Xu, Z., & Dwarkadas, S. (2003). *Peer-to-peer information retrieval using self-organizing semantic overlay networks*. Paper presented at the Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications, Karlsruhe, Germany.

Veropoulos, K., Campbell, C., & Cristianini, N. (1999). *Controlling the sensitivity of support vector machines.* Paper presented at the Proceedings of the international joint conference on artificial intelligence.

Veropoulos, K., Cristianini, N., & Campbell, C. (1999). The application of support vector machines to medical decision support: a case study. *Advanced Course in Artificial Intelligence*, 1-6.

Wang, S., Zhu, W., & Liang, Z.-P. (2001). *Shape deformation: SVM regression and application to medical image segmentation.* Paper presented at the Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on.

Welch, G., & Bishop, G. (1995). An introduction to the Kalman filter.

Yi-Lin, L., & Gang, W. (2005, 18-21 Aug. 2005). *Speech emotion recognition based on HMM and SVM.* Paper presented at the Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on.

Yu, S.-Z., & Kobayashi, H. (2003). An efficient forward-backward algorithm for an explicit-duration hidden Markov model. *Signal Processing Letters, IEEE, 10*(1), 11-14.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31*(1), 39-58. doi: 10.1109/tpami.2008.52