Processing Methods for the Detection of Landmark Acoustic Cues

by

Belinda Shi

B.S. Computer Science and Engineering
Massachusetts Institute of Technology 2021

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Feburary 2022

Author .....
Department of Electrical Engineering and Computer Science
Jan 14, 2022

Certified by .....
Stefanie Shattuck-Hufnagel
Principal Research Scientist, Research Laboratory of Electronics, MIT
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Processing Methods for the Detection of Landmark Acoustic Cues

by

Belinda Shi

Submitted to the Department of Electrical Engineering and Computer Science
on Jan 14, 2022, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

This paper presents work on an aspect of a new speech analysis system for lexical access, which is based on the concept of individual acoustic cues in the speech signal such as Landmarks, which are abrupt changes in the spectrum due to articulatory events associated with vowels and consonants. It provides an organized process that can easily be repeated and modified to be able to create an accurate and efficient detection module for landmark cues in speech files. The paper begins by examining patterns in the speech signal that may indicate the presence of vowel landmark cues, before proposing an algorithm that can predict the locations of vowel landmarks based on these observations. Then, it maps out a generalized system of steps needed to construct modules for detecting landmark acoustic cues, which involves extracting speech related measurements, processing them to accentuate certain characteristics, then using both speech production knowledge and mathematical analysis to determine which measurements are good indicators of certain acoustic cues. Finally, Gaussian Mixture Models using the selected raw and processed measurements are trained in order to efficiently and accurately distinguish landmark cues. These steps are applied to Vowel and Glide landmarks to develop a module that can distinguish them from other landmark cues in a speech signal. Development of this module provides a critical step in the development of a cue-based speech recognition system which can model human speech perception.

Thesis Supervisor: Stefanie Shattuck-Hufnagel
Title: Principal Research Scientist, Research Laboratory of Electronics, MIT

Acknowledgments

I would like to extend my sincere thanks to my supervisors, Stefanie Shattuck-Hufnagel and Jeung-Yoon Elizabeth Choi, for their consistent support and guidance during my research and the writing of this paper. This would not have been possible without your expert advice and continuous encouragement.

Contents

List of Figures

Chapter 1

Introduction

Speech recognition is used across many aspects of modern daily life. People carry smartphones with virtual assistants that respond to their spoken requests, video sites use voice recognition software to automatically generate transcriptions, and accessibility features to accommodate for disabilities use it extensively. Today's best speech recognition systems rely on machine learning techniques such as neural nets to achieve low error rates [6]. However, these approaches fall far short of human speech recognition performance, and also require heavy computational power, large sets of training data, and extensive training time. From a linguistics research perspective, these deep learning structures are opaque and do not reveal insights about the processes in a human brain that interpret speech. This motivates the search for an alternative method of characterizing the speech signal that is computationally light and more closely approximates the process of human speech perception.

In his 2002 paper [11], Stevens postulated that the human brain does not recognize speech by matching sound waves directly to words, but rather that listeners first naturally recognize phonologically relevant acoustic patterns known as acoustic cues in the speech signal. These acoustic cues provide evidence for identifying distinctive features, which are then used to discern word sequences. Current speech recognition systems, such as Deep Speech 1[6], do not make use of acoustic cues, and thus are not simulating the manner in which human listeners perceive speech. Working on the basis of Stevens' research, the MIT Speech Communication group [2] has proposed a

hierarchical speech recognition system that analyzes speech starting from the signal to the acoustic cue and to the lexical level that more closely approximates the human speech analysis process. It is composed of various modules that each detect a different acoustic cue in the signal, and uses the output of all of the modules to determine the distinctive features of the phonemes and words in a spoken utterance. Since all words are composed of phonemes which are described by a certain set of features, a simple matching algorithm can relate the features from the modules to words and sentences. This approach is expected to perform well in recognizing speech even where there is variation or modification, because it breaks down the speech signal into basic linguistic components. Thus, the system should generalize well to different accents, genders, and languages other than English.

1.1 Definitions

This section defines terms used throughout this paper and provides an overview of relevant speech production and acoustic phonetic theory concepts.

1.1.1 Human Speech Production

The source-filter model of human speech production is made up of several components:

1. Lungs (energy source)
2. Vocal folds/noise-producing structures (quasi-periodic/aspiration/frication source)
3. Vocal tract (resonance structure)

In order to produce a sound, a human speaker pushes air out of their lungs, which introduces **Energy** to the system. This force of air then travels through the vocal folds in the larynx. The amount of energy (or the **amplitude** of the sound wave) is related to the perceived loudness of the sound produced. A speaker can vibrate the vocal folds to produce a quasi-periodic sound source that can be characterized by a spectral representation with a harmonic structure. This creates a sonorant sound,

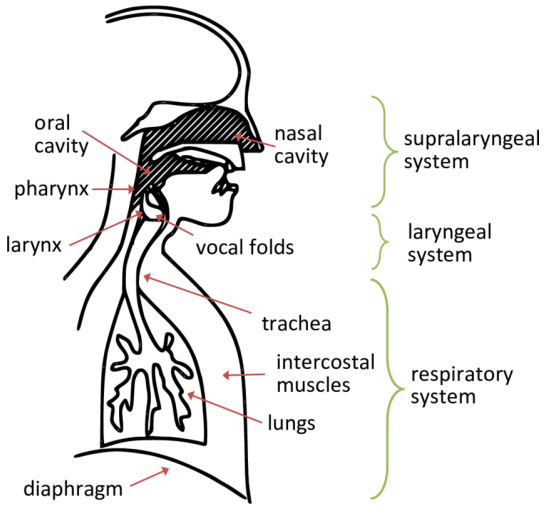

Figure 1-1: The human speech production system [7]

and there are two non-sonorant sources that can also be produced: aspiration and frication. The sound wave created by this source then passes through the articulatory tract which acts as a resonance filter and modulates the sound. This tract includes the passage to the lips (oral tract) and can also be coupled to the nostrils (nasal tract). Depending on the configuration of these articulators, we can produce different speech sounds related to their underlying phonemes.

1.1.2 Fundamental Frequency, Formants, Amplitude, and Energy Bands

Vibration of the vocal folds produces a quasi-periodic sound source with a spectral representation that is composed of the fundamental frequency, or F0, and harmonics at multiples of the fundamental frequency. This spectrum is the Source Spectrum in Figure 1-2. This source is multiplied by the filter created by the vocal tract to form the output spectrum, which is the resulting sound produced.

The filter applied to the sound source is described by formants, which are the resonant frequencies of energy resulting from particular configurations of an open vocal tract. They are numbered consecutively upwards from the lowest frequency, starting with F1, which provides information about the height of the tongue, and F2,

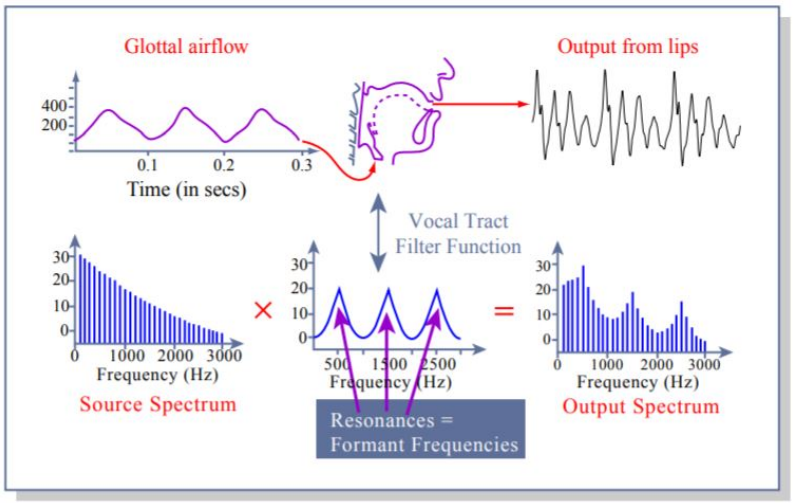

Figure 1-2: Source-Filter Model of Speech Production [3]

which relates to the frontness/backness of the tongue. There are usually on average five relevant formants that can be observed in human speech. Because the frequencies of the formants change based on the shape and position of the vocal tract, they can be used to determine what sound is being produced. Figure 1-3 shows F1 to F5 for the vowel sound in the word 'fog' plotted as horizontal red dotted lines. At the point in time marked by the vertical red dotted line, the value of F1 is 749 Hz.

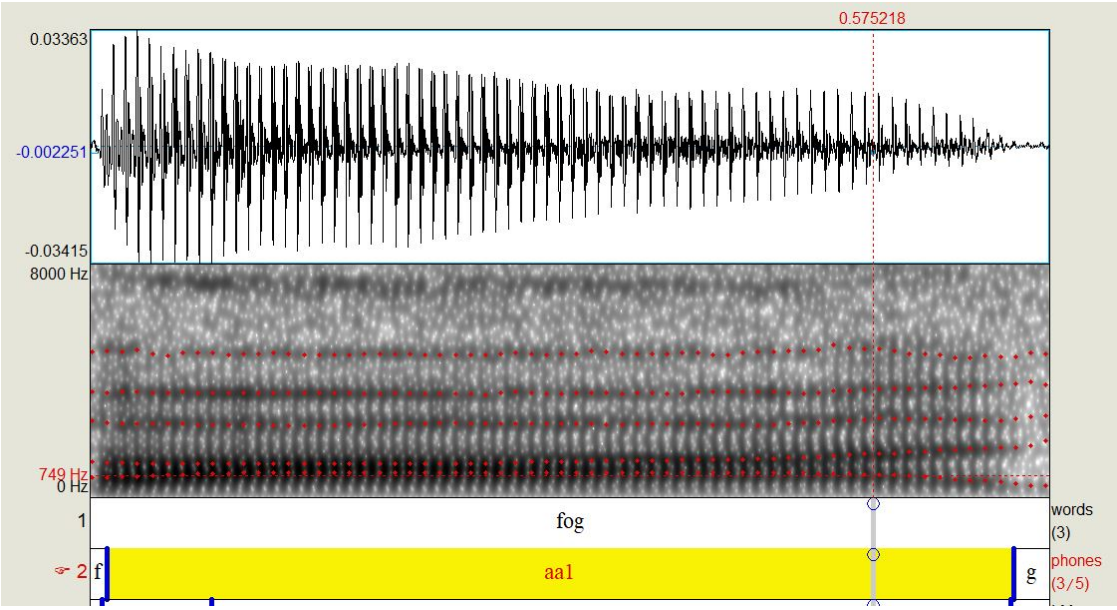

Figure 1-3: The formants of the vowel sound in 'fog' on a spectrogram

Another relevant speech measurement used in this paper is the amplitude of the energy at the formant frequencies. These are labelled with the same number as the corresponding formants, so A1 is the magnitude of F1 in a speech signal, A2 is the magnitude of F2, and so on.

The final type of measurement used in this study is found from energy bands, which separate the frequency spectrum into buckets and record the total amplitude of energy in each. Each band is roughly 1000Hz wide, and 8 bands are extracted, up to a maximum frequency of 8000Hz, labelled as E1, E2, E3 etc. The ranges of the energy bands are described in Figure 1-4, and are visualized on the frequency spectrum in Figure 1-5.

Energy Band	Start (Hz)	End(Hz)
E1	250	999
E2	1000	1999
E3	2000	2999
E4	3000	3999
E5	4000	4999
E6	5000	5999
E7	6000	6999
E8	7000	7999

Figure 1-4: Energy band buckets

1.1.3 Speech Landmarks

Landmarks are acoustic cues that are correlated with changes in speech articulation. There are four main categories of speech landmarks: vowels (V), glides (G), consonant closures (Cc), and consonant releases (Cr). Vowels and glides are marked at the maximum and minimum (respectively) of F1 (often approximated by amplitude) of the signal, while consonants are marked with a closure landmark at the start and a release landmark at the end of the relevant interval. The consonant landmarks are further separated into stops (S), fricatives (F), and nasals (N). Identifying landmark locations allows for categorisation of similar speech patterns into groups for analysis and prediction, which is essential for identifying words from the acoustic signal.

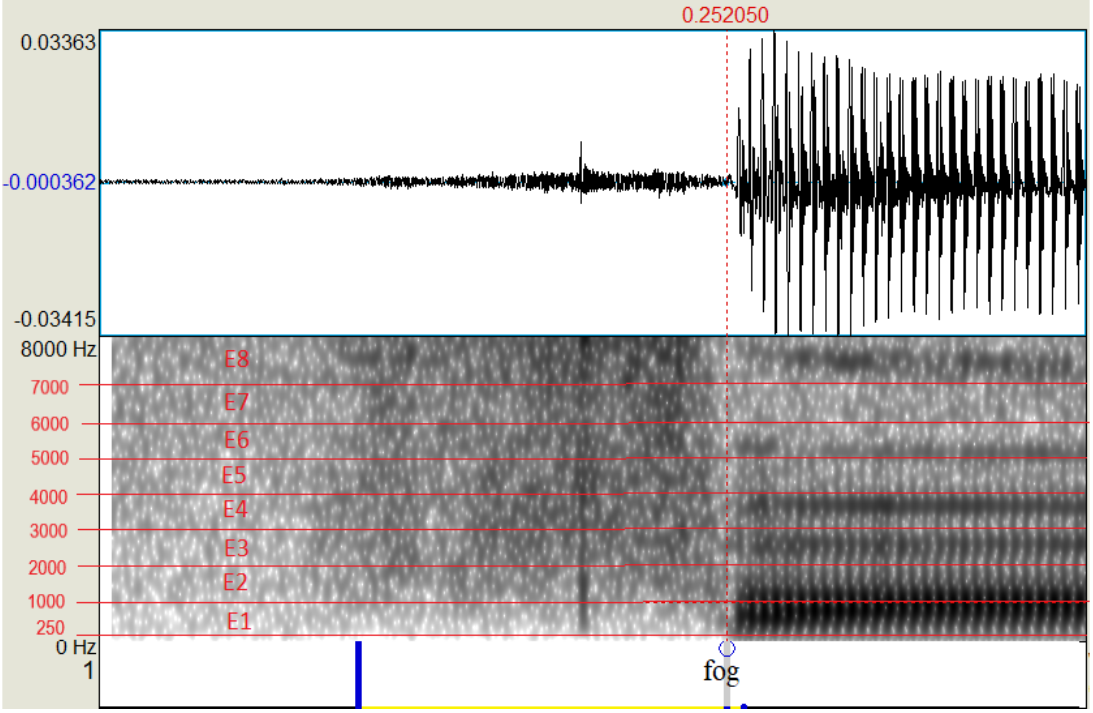

Figure 1-5: Energy bands visualized on the frequency spectrum

Landmark Type		General Label	Specific Label
Vowel		V	V
Glide		G	w, y, r, l, h
Consonant Closure	Stop	Sc	b-cl, d-cl, g-cl, p-cl, t-cl, k-cl, jh-cl, ch-cl, dj-cl
	Fricative	Fc	v-cl, dh-cl, z-cl, zh-cl, f-cl, th-cl, s-cl, sh-cl, jh-1, ch-1, dj-1
	Nasal	Nc	m-cl, n-cl, ng-cl
Consonant Release	Stop	Sr	b, d, g, p, t, k, jh-1, ch-1, dj-1
	Fricative	Fr	v, dh, z, zh, f, th, s, sh, jh-2, ch-2
	Nasal	Nr	m, n, ng

Figure 1-6: Types of landmarks and their corresponding labels

1.2 Research Goals

The overall goal of this research is to identify and organize the processing steps of extracting landmark acoustic cues to make modules for the overall hierarchical system. These are the 8 general labels in Figure 1-6; note that the final column contains more specific labels, which were in the past used for labelling speech files. The work described in this paper is focused primarily on detecting the Vowel and Glide landmark cues and separating them from the Consonant landmark cues.

The paper begins by examining patterns in the speech signal that may indicate

the presence of Vowel landmark cues, before proposing an algorithm that can predict the locations of Vowel landmarks in a speech file based on these observations. Then, it maps out generalized steps that are required to produce a landmark cue detection module. This involves producing a standardized system for extracting speech related measurements and filtering them to accentuate certain characteristics. Next, both speech production knowledge and mathematical analysis are used to determine which measurements are good indicators of certain acoustic cues. Finally, Gaussian Mixture Models, which are chosen for their fast and flexible modeling of clustered data, are trained using the selected raw and filtered measurements in order to accurately distinguish Vowel and Glide landmark cues from the consonant landmark cues efficiently. The steps taken to create the Vowel and Glide landmark detection module are generalized to result in an organized process that can easily be repeated and modified to be able to create an accurate and efficient detection module for any landmark cue in speech files.

These landmark detection modules play an essential part of the overall hierarchical speech recognition system. By following this process for each type of landmark, multiple models each detecting a different type of landmark cue can be created and run in a parallel fashion to detect all the landmarks in a given speech sample. These landmarks can then be used to identify distinctive features, which is an essential step towards recognizing words in the speech signal. This framework for speech analysis is more consistent with Stevens' [11] model of human speech perception.

Chapter 2

Data and Speech Related Measurements

2.1 Datasets

The datasets used in this paper include the TIMIT Acoustic-Phonetic Continuous Speech Corpus [5], and the Isolated Words dataset developed at the University of Connecticut, Storrs.

2.1.1 TIMIT

The TIMIT Acoustic-Phonetic Continuous Speech Corpus [5] features 630 American English speakers of eight major dialect regions reading sentences in WAV format. The sentences spoken are phonetically rich, which means they use a wide range of speech sounds for maximum phonetic coverage. 40 of these files featuring 2 male and 2 female speakers were chosen and labelled by trained MIT student researchers for this project.

2.1.2 Isolated Words

The Isolated Words dataset is developed by the University of Connecticut and features American English speakers, 3 female and 1 male, speaking about 1200 monosyllabic

	Example Sentences in TIMIT
1	She had your dark suit in greasy wash water all year.
2	Don't ask me to carry an oily rag like that.
3	Fill small hole in bowl with clay.
4	Assume, for example, a situation where a farm has a packing shed and fields.
5	I'll borrow some money from someone and go home by bus.
6	Biblical scholars argue history.
7	Publicity and notoriety go hand in hand.
8	You always come up with pathological examples.
9	Those answers will be straightforward if you think them through carefully first.
10	We'll serve rhubarb pie after Rachel's talk.

Figure 2-1: Sample of phonetically rich sentences in TIMIT in the sampled files.

words. A subset of these files (about 1200 words from one female speaker) were selected for this project and labelled by trained MIT researchers.

2.1.3 Labelling

The datasets are labelled and checked by students researchers in the Speech Communication Group at MIT. The main tool used is Praat, a software package for phonetic speech analysis designed by Boersma and Weenink [1] in the University of Amsterdam in 1991. The latest release is version 6.1, which was released in July 2019.

The Praat editor window shows the waveform of the sound file, followed by the a spectrogram showing the spectral characteristics of the sound over time at the top of the screen. Below, researchers can add Tiers, each for a different type of speech units to be labelled in the file. There are two types of Tiers, interval tiers and point tiers. Interval tiers label a range of time in the file, and is used for the words and phones/phonemes tiers, which are the first two tiers in Figure 2-2. The following tiers are for each broad type of acoustic cue: landmarks, vowel and glide place, consonant place, nasal, and glottal. Since this paper focuses on landmark cue detection, it only makes use of the landmark tier, which is named 'LM' and is the fourth tier in Figure 2-2. Researchers manually add landmark labels to this tier by inspecting both the waveform and spectrogram to determine where the labels should be placed, as well as listening to the sound. Vowel landmarks are placed at the maximum of the sound wave in the time interval where the formant structure is observed, and Glide

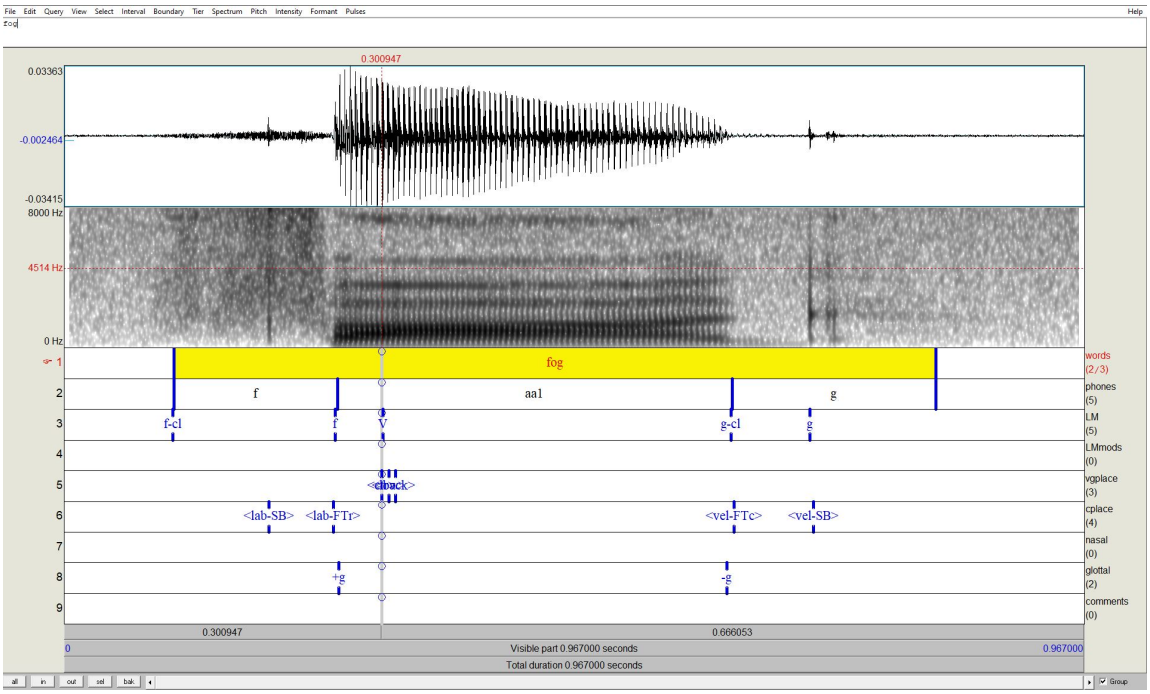

Figure 2-2: Praat Editor Window

landmarks are placed at the minimum. Consonant closures and releases are placed at the abrupt onsets and offsets of the vocal tract constriction intervals. The labels are placed by one researcher and checked for accuracy by at least one other researcher. For previously labeled files, the Specific Labels from Figure 1-6 are used, although this paper only considers the landmarks grouped by their General Label, which is the direction that the current labelling scheme is adopting for future work.

2.2 Speech Related Measurements (SRMs)

Matlab and Praat are used in order to extract measurements from the files in WAV format in the datasets. Each measurement is extracted by a different Matlab script. For all measurements, this paper uses WAV files with a sample rate of 44100 Hz, a window size of 20 frames and a frame shift of 1. This means if the WAV file is s seconds long, each measurement records $44100 \times s - 19$ samples.

The script for extracting energy band values uses the `spectrogram` function in the Matlab Signal Processing Toolbox, sums up the spectral magnitude in each of

the frequency ranges corresponding to each energy band as shown in Figure 1-5, and finally normalizes the values to be between 0 and 1. The scripts for extracting the formant values calls Praat scripts that use Praat's formant finding functions. There is one script for finding F0 (or Pitch) and another for finding the rest of the formants. Finally, the last script gets the spectral magnitude of the signal for each of the formats, which are the amplitude measurements. These measurements are combined into a 2-dimensional array where each column is a different measurement. A dataset is represented as a struct with the name of each WAV file as a fieldname, and the 2D SRM array above under the 'SRM' field in each.

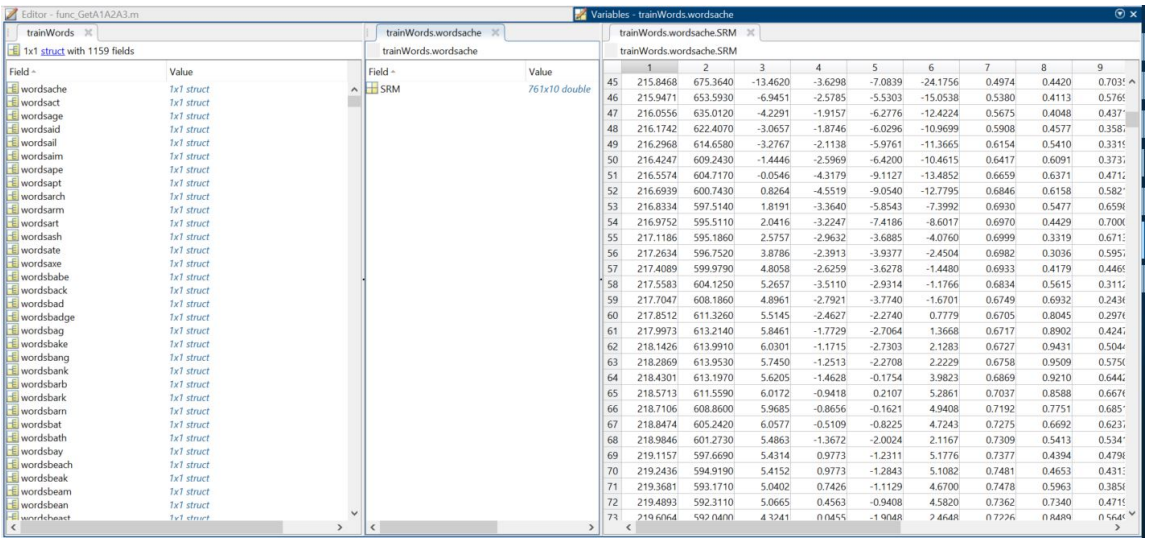

Figure 2-3: Structure of dataset SRM struct. The left window shows all the filenames in the dataset. The center window shows the singular SRM fieldname in the data for each file. The right window shows the 2D array where each column is a different measurement.

Chapter 3

Detecting Vowel Landmarks

This section proposes a vowel recognition module that uses knowledge about the acoustic qualities of human speech production to identify vowels based on their characteristics. It shows that vowel landmarks can be found by analyzing patterns in changes of speech measurements over time, and an algorithm based on these observations can be both accurate and efficient.

3.1 Detecting Vowels From Properties

Since vowels are usually voiced, vowel landmarks can be found in regions where there is a harmonic structure, and therefore a fundamental frequency (F0). In Figure 3-1, each red dotted line is a vowel landmark marked at that time in milliseconds, and F0 (in Hz) is marked in blue. The regions with no corresponding F0 value are regions where the speaker is not producing a vocal fold vibration. Given F0, we can extract ranges of samples in the speech signal where vowels are most likely to appear.

Plotting the value of the first three energy bands, it is clear that vowels appear near the large peaks. However, it seems that no particular energy band peaks at every vowel. For example, in Figure 3-2, the first two vowels correspond with large spikes in E1 and E2 respectively, and the other energy bands do see a small increase, but not as much. This suggests that the sum of E1, E2, and E3 may be a good indicator of a vowel landmark.

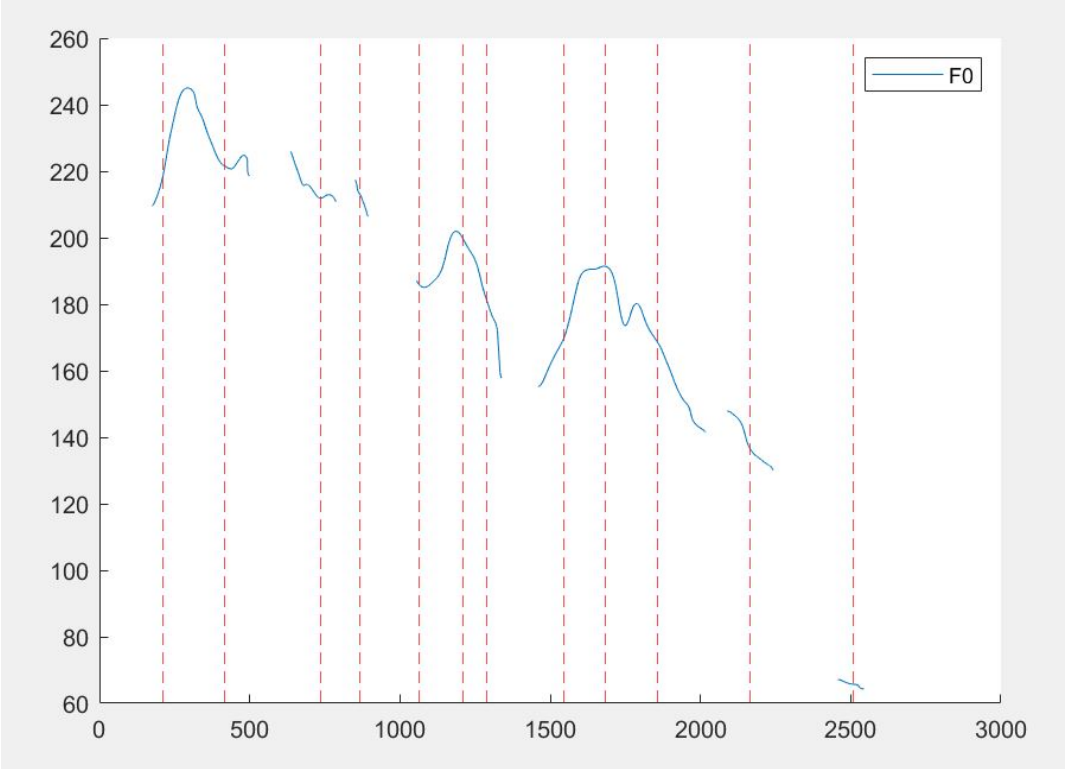

Figure 3-1: Plot of F0 with labelled vowel landmarks as red dotted lines

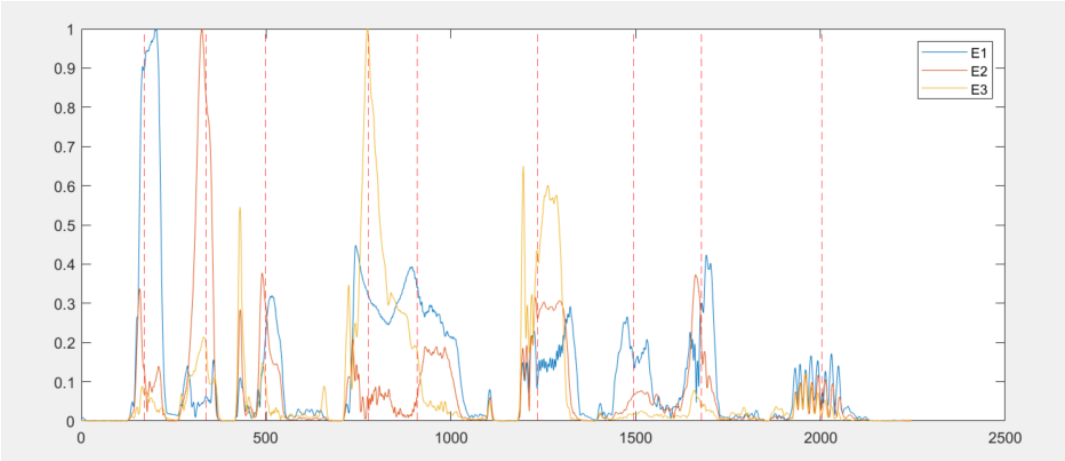

Figure 3-2: Plot of E1, E2, E3 with labelled vowel landmarks as red dotted lines

After finding the approximate locations of vowel landmarks, we can more precisely locate where the vowel label should be placed by looking at the values of A1, A2, and A3. The labels should be placed at the maximum amplitude of the vowel, so they should appear at the peak of these measurements. The manually labelled vowels follow this pattern, as can be seen in Figure 3-3.

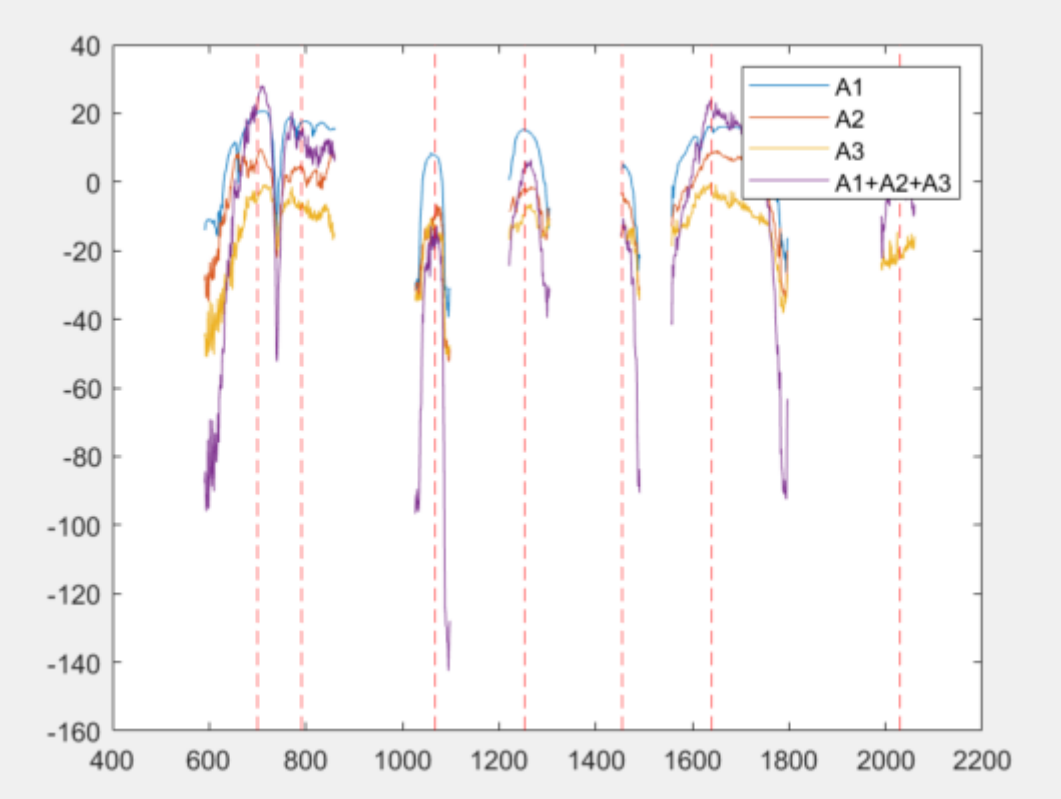

Figure 3-3: Plot of A1, A2, and A3 with labelled vowel landmarks as red dotted lines

3.2 The Vowel Landmark Module

From these speech related measurements, the proposed algorithm is as follows:

1. Find ranges where vowels are most likely to exist from the segments where F0 has a value (i.e the value of F0 is not NaN). This is visualized in Figure 3-4 with the blue line.

2. For each of these ranges, find the all the local maxima of a smoothed E1+E2+E3 within them. This measurement is smoothed using the matlab `smoothdata` function, which returns a moving average of the input using a heuristically determined fixed window length. This reduces the effect of noise in the data. The resulting points are approximations of where vowels are in the speech file. These points are represented as green stars in Figure 3-4.

3. For each maxima found, find the closest maxima of A1+A2+A3 (blue stars)

and place a vowel label in that location. Maxima are found with a minimum peak distance of 100 frames to pick only the significant peaks and exclude small peaks due to noise. This moves the point to the most accurate location, as can be seen comparing these points (red stars) to the actual vowel locations (red lines) in Figure 3-4.

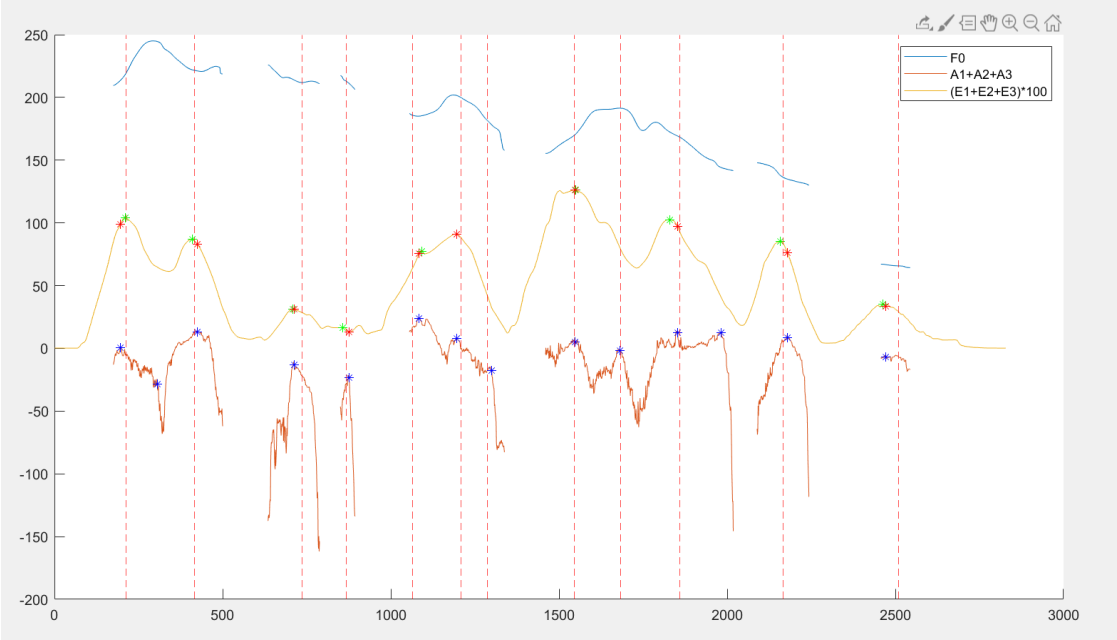

Figure 3-4: The algorithm visualized, where the green stars are the peaks in E1+E2+E3 and the red stars are those points shifted to the closest peak in A1+A1+A3 (blue stars). Red dotted lines are manually labelled vowel locations.

This method is most similar to how a trained researcher might recognize a vowel when manually labelling files by visually inspecting the speech signal. Accuracy is measured by comparing the resulting labels outputted by the algorithm to the manually marked vowel labels.

3.3 Results and Analysis

The effectiveness of the algorithm is analyzed by comparing the manually marked vowel locations for the files in the TIMIT database to the locations outputted by the module.

A point is classified as a True Positive (TP) if the module detects a vowel within a 40ms window around it. If the module does not output a vowel location within this window, the point is classified as a False Negative (FN). Furthermore, if the module outputs a vowel location and there is no labelled vowel within the specified 40 ms window size around that location, this is a False Positive (FP). These results can be seen in Figure 3-5.

Dataset	TP	FP	FN	Precision	Recall	F1
TIMIT (all)	342 88%	46 12%	103 27%	0.881443	0.768539	0.821128
TIMIT (male)	158 86%	25 14%	62 34%	0.863388	0.718182	0.784119
TIMIT (female)	184 90%	21 10%	41 20%	0.897561	0.817778	0.855814

Figure 3-5: Results on the selected subset of TIMIT files with acoustic cues labelled

The algorithm achieved a F_1 score of 0.82 on all of the TIMIT files, meaning it is quite accurate across diverse speakers. Comparing the separated F_1 scores of the male speakers and female speakers in this dataset reveals that the algorithm performs slightly better on the female speakers, but only by a small margin.

In comparison, the associative neural network method used in Gangashetty et al [4] results in 68% TP, 6.6% FP, and 32% FN. Compared to this method, the algorithm presented in this paper results in a slightly higher F_1 score, but a significantly higher Recall and a much lower precision (see figure 3-6). This means that while this algorithm is more accurate when detecting vowels, it also makes more mistakes classifying other acoustic cues as vowels.

Algorithm	TP	FP	FN	Precision	Recall	F1
Proposed	88%	12%	27%	0.881443	0.768539	0.821128
Gangashetty et al	68.19%	6.65%	31.8%	0.911146	0.681968	0.780072

Figure 3-6: Comparison of results

The algorithm is very fast and can run on limited resources. On a personal laptop with 16GB memory and Intel Core i7, it achieves about 1 second runtime to output vowel locations for 40 TIMIT audio files. Figure 3-7 shows expected runtime of the

algorithm as the number of files increases. As the algorithm works on one file at a time, the trend is expected to be linear, with a small constant offset for other work such as the initial function setup. It also doesn't require selection or training of models, but instead is based on theoretical understanding of the physics of speech production.

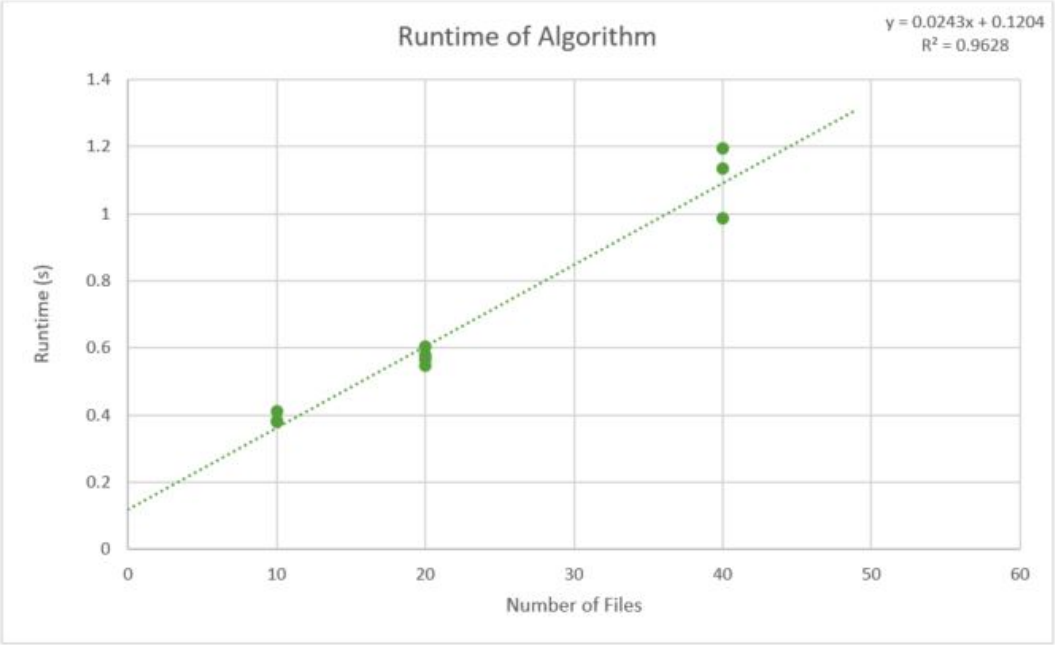

Figure 3-7: Expected Runtime

A limitation of this method may stem from human error in labelling. While a vowel is easily detectable by humans, it can span a long period of time, and it is difficult for a manual labeller to determine the location of the peak when there is noise, so they may place the label somewhere else in the vowel. If this difference is greater than the chosen window size, it may be counted as a FP or a FN. Increasing the window size would not solve this issue as doing so could match incorrect predictions of vowels in consonant locations to neighboring vowels, and count them as a correct prediction.

Overall, the proposed algorithm performs well and is able to be run on systems with low resources, so it will be a good fit as a module for the greater hierarchical system.

Chapter 4

Generalizing To All Landmarks

From the method used to build the vowel landmark detection module in the previous chapter, a generalized formula for building a detection module for any landmark cue can be theorized. First, since landmark cues are speech events that are signalled by a change of articulation characteristics, they correspond to changes in one or more of the speech related measurements. If we can find which measurements are most likely to be good indicators of a landmark cue being present, we can build a module that specifically detects that cue.

4.1 Filtering Measurements

For vowel landmark detection, the raw values of F0 were used to determine whether the sound is voiced. However, for the other two measurements (E123 and A123) that were used, the algorithm needed to find where those measurements peaked. Other landmarks may need to also find peaks, or even dips in values. Matlab scripts are used to apply a filter to the original SRM data structure to accentuate each of the following features in Figure 4-1. The output has the same structure as the original SRM representation in Figure 2-3, with each file as a fieldname and the data as a 2-dimensional array of values.

Filter	Description
Peaks	Prominence of the peak at peak locations, half of the prominence in a 30 ms window around the peak
Dips	Prominence of the dip at dip locations, half of the prominence in a 30 ms window around the dip
Voiced	0 if F0 is NaN, or 1 otherwise
DiffVowel	Each data point is the difference between its raw measurement value and the value of the closest vowel landmark to its location
Derivative	Derivative of the (smoothed) measurement, useful for detecting overall slopes and trends

Figure 4-1: List of Matlab data filters. Note that the filter Voiced is based on detected phonation, rather than whether the corresponding signal is related to the production of an underlying voiced phoneme.

4.2 ANOVA

Which measurements, or processed measurements, are the most useful for detecting which landmarks?

Past phonetics and speech production research have shown that certain landmarks have particular characteristics that can be seen in the speech related measurements taken at that landmark. This information is a good starting point for selecting measurements to build a module for a specific landmark cue, but another method may be used to check that the chosen measurements are strongly correlated with the locations of that landmark. This involves performing one-way analysis of variance (ANOVA) on the measurements against the locations of the landmarks. The ANOVA algorithm determines the effect of independent input variables on the output variable. Each measurement, as well as each processed (or filtered) measurement, is run against each of the groups of landmark cues. In the following table, a smaller p-value from the ANOVA algorithm (green) means that the measurement is more likely to be able to distinguish the target landmark from the other non-target landmarks. Note that the ANOVA algorithm assumes that the input variables are independent, which may not be the case for the speech data, but is sufficient for the purpose of determining

which measurements to choose for training Gaussian Mixture Models in the following chapter. The combined group of Vowel and Glide landmarks was included as an extra target, since they are characteristically similar, and could be detected together in the first step of a detection system that first separates consonantal and non-consonantal landmarks. From this analysis we are able to confirm that the labelled speech data shows trends and correlations that correspond with existing knowledge about how different landmarks are produced.

	G	V	G + V	Sc	Sr	Fc	Fr	Nc	Nr
F0	2.15E-01	1.70E-01	7.09E-02	7.95E-02	6.57E-01	4.42E-01	6.12E-02	5.13E-01	1.54E-01
F1	1.19E-10	3.07E-26	1.13E-44	2.61E-01	3.76E-13	2.57E-19	3.78E-20	1.60E-04	7.47E-01
A1	9.61E-04	8.98E-83	1.61E-114	8.86E-06	5.37E-05	1.26E-15	4.45E-03	2.54E-09	4.79E-12
A2	8.65E-08	2.94E-61	1.09E-94	2.33E-07	1.81E-02	4.21E-13	4.35E-03	1.58E-07	2.58E-08
A3	7.29E-12	9.61E-64	7.23E-78	4.24E-08	2.25E-01	1.16E-05	1.40E-03	1.48E-07	8.02E-04
A1+A2+A3	8.76E-05	9.58E-88	8.31E-117	1.81E-08	4.06E-03	9.42E-13	8.02E-04	1.38E-09	1.95E-09
E1	1.24E-11	1.40E-89	8.10E-126	1.88E-21	1.20E-09	7.15E-16	1.36E-03	3.31E-04	2.90E-01
E2	9.15E-07	7.63E-91	1.61E-112	1.82E-13	7.77E-05	1.94E-10	1.86E-02	2.55E-06	8.89E-02
E3	1.78E-05	4.06E-62	3.47E-75	8.71E-09	6.64E-02	2.25E-05	7.79E-03	2.92E-04	4.48E-01
E1+E2+E3	3.12E-12	1.77E-143	1.21E-195	1.07E-25	1.22E-08	5.78E-19	5.60E-05	9.69E-08	1.02E-01

Figure 4-2: P-values generated by ANOVA on unfiltered measurements. Columns are target landmarks and rows are different measurements

	G	V	G + V	Sc	Sr	Fc	Fr	Nc	Nr
dF0	8.18E-01	5.43E-01	7.72E-01	9.64E-01	1.00E+00	1.39E-03	1.00E+00	3.38E-01	4.33E-01
dF1	8.85E-01	6.43E-04	1.76E-05	1.12E-15	2.62E-04	4.22E-36	6.83E-26	1.60E-01	1.56E-04
dA1	1.77E-01	8.64E-02	1.08E-04	1.75E-06	7.47E-17	1.01E-12	9.39E-03	9.05E-04	3.00E-02
dA2	3.91E-03	1.38E-01	1.54E-07	3.01E-15	2.11E-15	2.79E-07	1.17E-04	5.00E-06	1.62E-04
dA3	1.92E-03	4.23E-02	3.85E-06	1.59E-15	4.31E-11	1.20E-10	9.21E-04	2.09E-05	1.09E-05
d(A1 +A2+A3)	7.20E-01	3.82E-02	6.67E-09	1.10E-16	1.07E-14	6.63E-15	1.21E-05	4.45E-07	2.49E-05
dE1	9.68E-06	2.68E-04	7.20E-09	5.37E-28	6.66E-17	2.90E-19	2.19E-20	2.73E-02	6.47E-04
dE2	5.21E-10	5.76E-05	1.44E-10	1.15E-18	1.60E-13	1.64E-03	1.10E-12	2.93E-04	2.23E-05
dE3	9.60E-08	2.23E-03	1.54E-06	6.39E-10	6.87E-06	7.12E-01	4.31E-03	2.50E-03	2.94E-05
d(E1+E2+E3)	4.05E-08	6.20E-06	4.00E-13	1.72E-30	2.30E-20	1.84E-10	8.18E-20	1.48E-04	3.37E-07

Figure 4-3: P-values generated by ANOVA on filtered (Derivative) measurements. Columns are target landmarks and rows are different measurements

	G	V	G + V	Sc	Sr	Fc	Fr	Nc	Nr
Peaks F0	7.26E-01	4.24E-01	6.56E-01	4.12E-01	2.50E-01	9.78E-01	9.74E-01	6.47E-04	8.50E-01
Peaks F1	2.76E-08	2.19E-32	2.58E-49	2.75E-14	2.33E-12	1.18E-05	3.51E-06	1.88E-02	9.27E-01
Peaks A1	6.28E-01	1.38E-52	1.25E-50	1.06E-01	1.50E-02	1.40E-02	2.32E-01	5.05E-03	4.24E-03
Peaks A2	8.86E-02	1.11E-40	5.00E-43	4.08E-02	6.72E-03	5.82E-02	9.31E-02	6.01E-03	2.23E-03
Peaks A3	1.21E-01	5.64E-46	1.22E-43	3.52E-02	1.30E-02	6.98E-02	3.70E-01	1.22E-02	1.14E-01
Peaks A1+A2+A3	5.66E-02	7.43E-58	1.05E-56	2.72E-02	1.48E-02	2.96E-02	1.44E-01	4.61E-03	1.54E-03
Peaks E1	4.29E-02	1.06E-67	6.47E-71	3.23E-05	2.68E-05	1.63E-04	1.07E-01	1.52E-02	5.76E-02
Peaks E2	1.01E-01	8.51E-73	1.09E-74	8.80E-05	4.92E-04	8.14E-04	8.60E-02	3.89E-02	1.17E-02
Peaks E3	2.02E-01	8.47E-55	2.90E-55	3.90E-03	3.85E-02	1.22E-02	4.96E-02	3.50E-02	7.03E-02
Peaks E1+E2+E3	1.58E-01	1.13E-108	7.63E-112	6.09E-06	1.01E-05	9.59E-05	4.06E-04	6.33E-03	4.33E-03

Figure 4-4: P-values generated by ANOVA on filtered (Peaks) measurements. Columns are target landmarks and rows are different measurements

	G	V	G + V	Sc	Sr	Fc	Fr	Nc	Nr
Dips F0	1.55E-01	9.06E-01	2.37E-01	9.91E-01	5.81E-03	9.65E-01	1.10E-02	4.01E-01	9.69E-02
Dips F1	8.85E-04	1.87E-20	2.12E-24	8.29E-13	1.11E-06	4.88E-03	6.92E-01	3.09E-01	1.91E-02
Dips A1	2.88E-04	1.59E-10	1.50E-12	2.42E-02	5.20E-12	1.33E-01	8.31E-01	8.26E-01	1.41E-09
Dips A2	5.99E-03	1.86E-11	2.01E-13	1.50E-03	4.85E-14	3.02E-01	9.19E-01	4.95E-01	1.02E-13
Dips A3	1.12E-06	7.05E-14	4.39E-17	6.28E-02	1.09E-09	2.64E-01	3.14E-01	9.06E-01	6.97E-07
Dips A1+A2+A3	5.65E-04	6.61E-11	1.63E-13	4.15E-04	7.28E-20	1.67E-01	8.15E-01	4.15E-01	2.07E-09
Dips E1	2.66E-04	1.67E-18	6.01E-25	1.26E-09	3.71E-08	8.87E-01	1.54E-02	9.59E-01	4.40E-02
Dips E2	8.34E-03	2.25E-17	9.68E-21	2.74E-13	7.98E-02	5.07E-02	4.15E-02	8.75E-01	1.95E-02
Dips E3	3.73E-02	3.64E-09	3.53E-11	3.50E-15	9.48E-01	9.10E-01	1.13E-01	6.93E-01	4.08E-01
Dips E1+E2+E3	2.16E-03	2.02E-19	4.47E-25	2.11E-08	4.97E-08	7.85E-01	4.82E-06	8.73E-01	5.21E-01

Figure 4-5: P-values generated by ANOVA on filtered (Dips) measurements. Columns are target landmarks and rows are different measurements

	G	V	G + V	Sc	Sr	Fc	Fr	Nc	Nr
DiffV F0	9.81E-01	1.38E-01	8.62E-01	3.32E-01	9.40E-01	1.58E-05	1.02E-03	1.11E-01	5.89E-01
DiffV F1	3.52E-13	6.91E-35	2.56E-60	2.20E-01	1.67E-18	3.45E-21	2.60E-34	3.41E-06	1.01E-01
DiffV A1	1.66E-01	5.81E-55	6.42E-58	4.94E-02	4.04E-18	1.04E-04	4.03E-01	5.29E-04	2.24E-07
DiffV A2	2.11E-02	5.58E-51	1.04E-58	2.40E-03	1.07E-13	5.86E-06	8.03E-01	9.86E-05	1.36E-11
DiffV A3	8.28E-03	3.14E-57	1.92E-58	1.65E-04	1.20E-09	1.99E-03	8.33E-02	2.73E-03	1.03E-09
DiffV A1+A2+A3	4.20E-01	2.02E-60	1.05E-63	7.17E-03	5.52E-20	6.63E-06	3.78E-01	4.81E-03	5.04E-10
DiffV E1	1.93E-05	3.78E-75	5.36E-93	1.22E-06	2.19E-11	8.18E-26	7.96E-07	4.36E-01	8.14E-01
DiffV E2	5.51E-02	9.38E-64	4.20E-69	1.22E-07	2.29E-09	6.28E-13	7.48E-02	9.07E-02	5.62E-01
DiffV E3	8.70E-01	1.70E-36	1.82E-33	1.71E-06	3.16E-04	1.05E-07	4.34E-01	4.73E-01	3.44E-01
DiffV E1+E2+E3	3.31E-02	9.36E-85	1.11E-94	1.45E-07	1.69E-13	1.75E-22	1.55E-04	1.87E-01	5.48E-01

Figure 4-6: P-values generated by ANOVA on filtered (Difference to nearest Vowel) measurements. Columns are target landmarks and rows are different measurements

	G	V	G + V	Sc	Sr	Fc	Fr	Nc	Nr
Voiced	1.04E-11	3.33E-51	1.76E-77	1.77E-01	8.94E-49	2.16E-06	2.24E-24	3.41E-04	3.83E-01

Figure 4-7: P-values generated by ANOVA on filtered (Voiced) measurements. Columns are target landmarks.

Chapter 5

Gaussian Mixture Models

Gaussian Mixture Models (GMM) are probabilistic models that classify data points generated from one or more Gaussian distributions. It is one of the most effective representations of clustered data, and is both fast and flexible. Since different acoustic cues have certain characteristics that can be seen in the speech related measurements data, the points representing each cue are likely to be clustered in one or more groups in the data space. We selected the GMM framework for constructing detection modules for acoustic cues, to compare with and/or augment the initial acoustic theory based method for detecting Vowel Landmark cues (Chapter 3). This method can also be easily applied to the detection of other acoustic cues.

Gaussian Mixture Models are trained using an iterative expectation maximisation algorithm, `fitgmdist` in Matlab. This algorithm starts with initial values for each Gaussian component's mean, covariance matrix, and mixing proportions, and iteratively maximises probabilities that each given data point is from the component corresponding to its label. It iterates until the values converge, or it reaches the maximum number of iterations, and returns the resulting GMM, which is fitted to the given data.

In order to successfully train a GMM, we need to be selective about the number of measurements used to train the model. Each measurement considered adds a dimension to the training space, and too many dimensions can cause the algorithm to either not converge or be overfitted to the training data. To avoid this, heuristically,

the number of measurements to select needs to be at most half the number of available target data points in the training set.

5.1 Vowel landmarks

In Chapter 3, the analysis on the characteristics of Vowel landmarks showed that three measurements were important in detecting them: $F0$, $E1+E1+E3$, and $A1+A2+A3$. The ANOVA results also support these observations, since they all achieved very low p-values in Figure 4-2. The next step is to use these measurements to train a GMM to detect Vowel landmarks. The GMM training algorithm is configured to to fit one Gaussian distribution to the set of all Vowel landmark cues, and one Gaussian distribution to the set of all other landmark cues, Using this 1 cluster model, the results of this detection scheme using TIMIT as training data and Isolated Words as testing data is in Figure 5-1. It is difficult to compare this GMM method to the algorithmic method of detecting vowel landmarks discussed in Chapter 3 because the previous method did not produce True Negatives, so accuracy cannot be compared. However this GMM method results in higher recall, but lower precision on average.

#points	V	GFSN
Model 1 (V)	656	478
Model 2 (GFSN)	20	2317

% total	V	GFSN
Model 1 (V)	18.8995%	13.7712%
Model 2 (GFSN)	0.5762%	66.7531%

Accuracy	0.8565
Recall	0.9704
Precision	0.5785
F1	0.7249

Figure 5-1: Results of training Vowel landmark GMM using raw measurements

Since the algorithm from the previous chapter used the peaks of E1+E2+E3 and A1+A2+A3, and F0 only to detect whether a data point is phonated speech, we can try to add filtered measurements to train an improved vowel detection model.

Figure 5-2 shows the results for training a vowel landmark detection GMM using Peaks E1+E2+E3, Peaks A1+A2+A3, E1+E2+E3, A1+A2+A3, and Voiced filtered measurements. This GMM results in better accuracy and much better precision, but has a sightly lower recall, ultimately resulting in a better F1 score.

#points	V	GFSN
Model 1 (V)	618	306
Model 2 (GFSN)	46	767

% total	V	GFSN
Model 1 (V)	17.8047%	8.8159%
Model 2 (GFSN)	1.6710%	71.7084%

Accuracy	0.8951
Recall	0.9142
Precision	0.6688
F1	0.7725

Figure 5-2: Results of training Vowel landmark GMM using filtered measurements

5.2 Vowel and Glide landmarks

The results from one-way ANOVA for each of the types of landmarks and each of the filtered SRMs show which measurements are most likely to be good candidates to choose for training a Gaussian Mixture Model, since the p-values from the algorithm are a indicator of how correlated the measurement is to the target data points.

Since vowels and glides are acoustically similar, a model that detects both may be more successful than a model that only detects glide landmarks. From the ANOVA results in Figure 4-2, the raw measurements $F1$, to $E1 + E2 + E3$ all have very small (less than $1E - 40$ p-values for Vowel and Glide landmarks together. Furthermore, from Figure 4-7, the filtered F0 (voiced) measurement is also a good candidate for differentiating vowels and glides. Thus, these measurements were selected to train two GMMs: one that calculated the probability of a data point being a Vowel or Glide landmark, and one that gives the probability of a data point being one of the consonant landmarks (Fricatives, Stops, and Nasals).

Because the number of selected measurements is 10, each landmark cue is represented by a data point in 10-dimensional space, which is difficult to visualize. For ease of viewing, the figures below are plotted in 2-dimensional space, with $F1$ along the y-axis and $A1 + A2 + A3$ along the x-axis. Figure 5-3 shows the landmark cues in the training set (TIMIT) separated into the two groups as defined above. There are 632 Vowel and Glide landmark cues (red), and 835 Fricatives, Nasals, and Stop cues (blue) in the TIMIT dataset.

Figure 5-4 shows the landmark cues in the testing set separated into the two groups as defined above. There are 1122 Vowel and Glide landmark cues (red), and 2349 Fricatives, Nasals, and Stop cues (blue) in the Isolated Words dataset.

1-Cluster Models

For the first set of models, we use a training parameter of 1 cluster for each model. That is, we train two models, one for Vowel and Glide landmarks, and one for Fricatives, Stops and Nasals, where each model is made up of exactly one Gaussian distri-

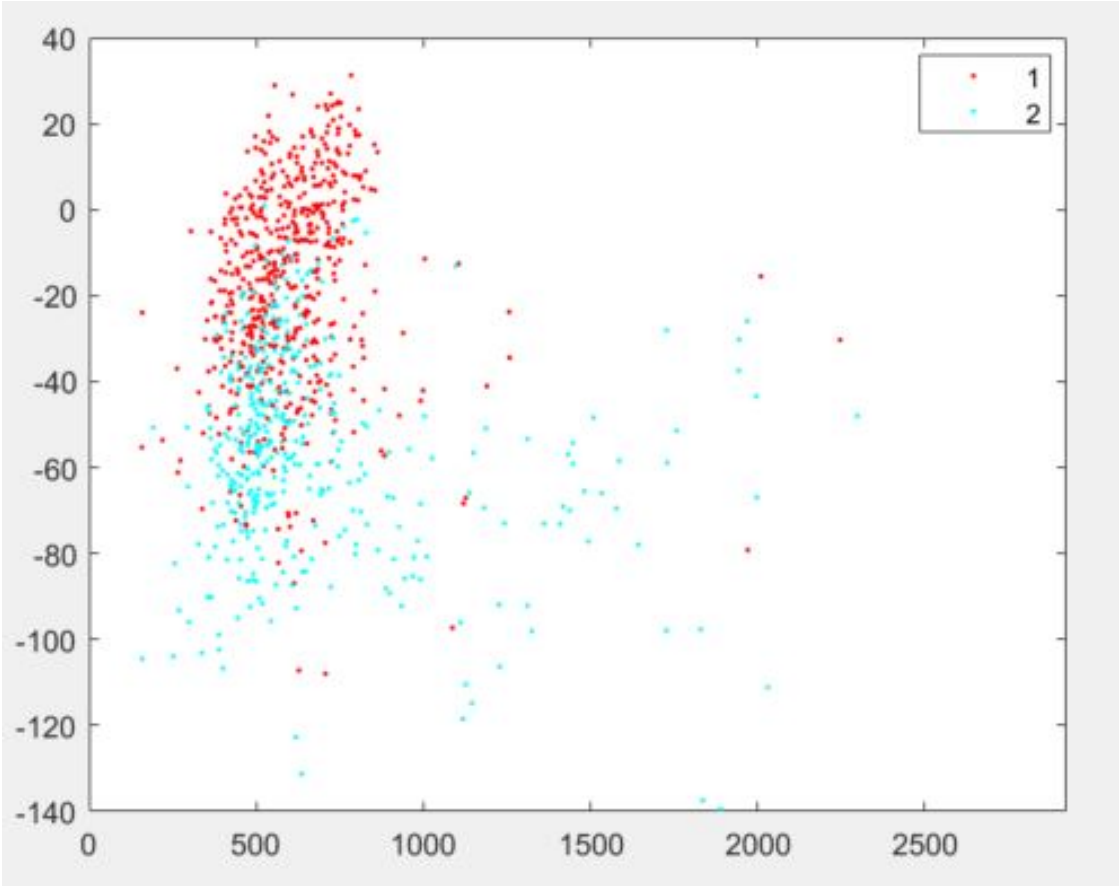

Figure 5-3: Landmark data points from TIMIT data in 2D space. Red dots are Vowel and Glide landmarks, and blue dots are Fricatives, Nasals and Stops

bution. This assumes that each of the two sets of data points are clustered around one point in the training space. The two models are trained on the TIMIT data, and tested on the Isolated words dataset. For testing, we calculate and compare the probabilities of each landmark cue in the testing set being in each of the models, and the point is classified according to the maximum probability score.

Figure 5-5 can be compared to Figure 5-4 to visually estimate its accuracy. The clusters show the same general shape, which suggests that the model was successful in classifying a large number of points. Furthermore, Figure 5-6 shows the distribution of the classified points as well as accuracy, precision, recall, and F1 score. Excluding the time it takes to read and process the files in the dataset, since this can be done in a pre-processing step and does not need to be repeated when training every new model, the Vowel and Glide landmark model required 3 iterations to converge to a log-

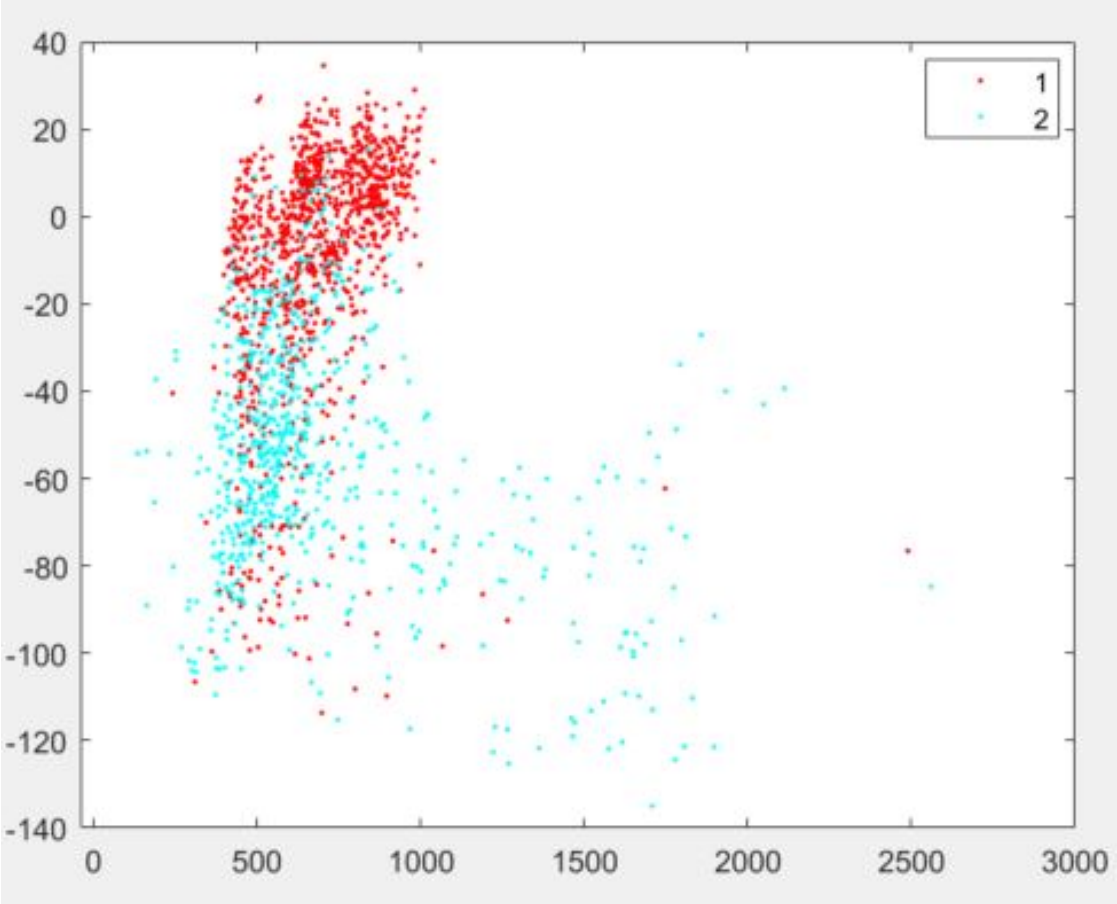

Figure 5-4: Landmark data points from the Isolated Words data in 2D space. Red dots are Vowel and Glide landmarks, and blue dots are Fricatives, Nasals and Stops

likelihood of -12252.1, which took 0.028589 seconds. The Fricative, Nasal, and Stop landmark model also required 3 iterations to achieve a log-likelihood of -7871.18, taking 0.101105 seconds. Thus, the total training time for the 1 cluster models is $0.028589 + 0.101105 = 0.129694$ seconds, and testing took 0.1121 seconds. Overall, the 1 cluster models achieved very high accuracy, recall, precision, and F1 score, and is very fast to train and use.

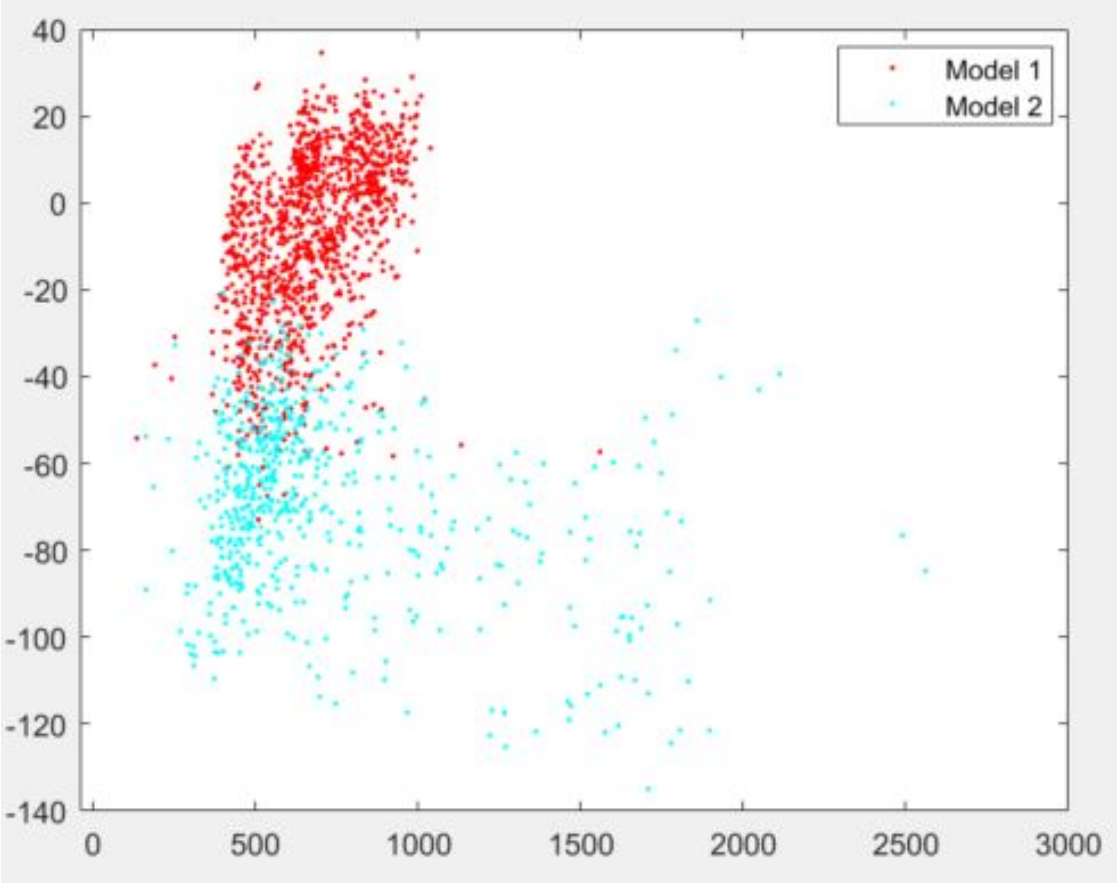

Figure 5-5: Classified points from Isolated Words data in 2D space for the 1-cluster models. Red dots are more likely to be from the GMM corresponding to Vowel and Glide landmarks, and blue dots are more likely to be from the GMM corresponding to Fricatives, Nasals and Stops

#points	VG	FSN
Model 1 (VG)	953	228
Model 2 (FSN)	169	2121

% total	VG	FSN
Model 1 (VG)	27.4561%	6.5687%
Model 2 (FSN)	4.8689%	61.1063%

Accuracy	0.8856
Recall	0.8494
Precision	0.8069
F1	0.8276

Figure 5-6: Results training GMMs with 1 cluster

2-Cluster Models

For the second set of models, we try a training parameter of 2 clusters for each model, which specifies that each model is made up of exactly two Gaussian distributions. This allows for a larger spread of data, since there can be two centers around which the points are clustered, and may more accurately model the distribution of landmark cues.

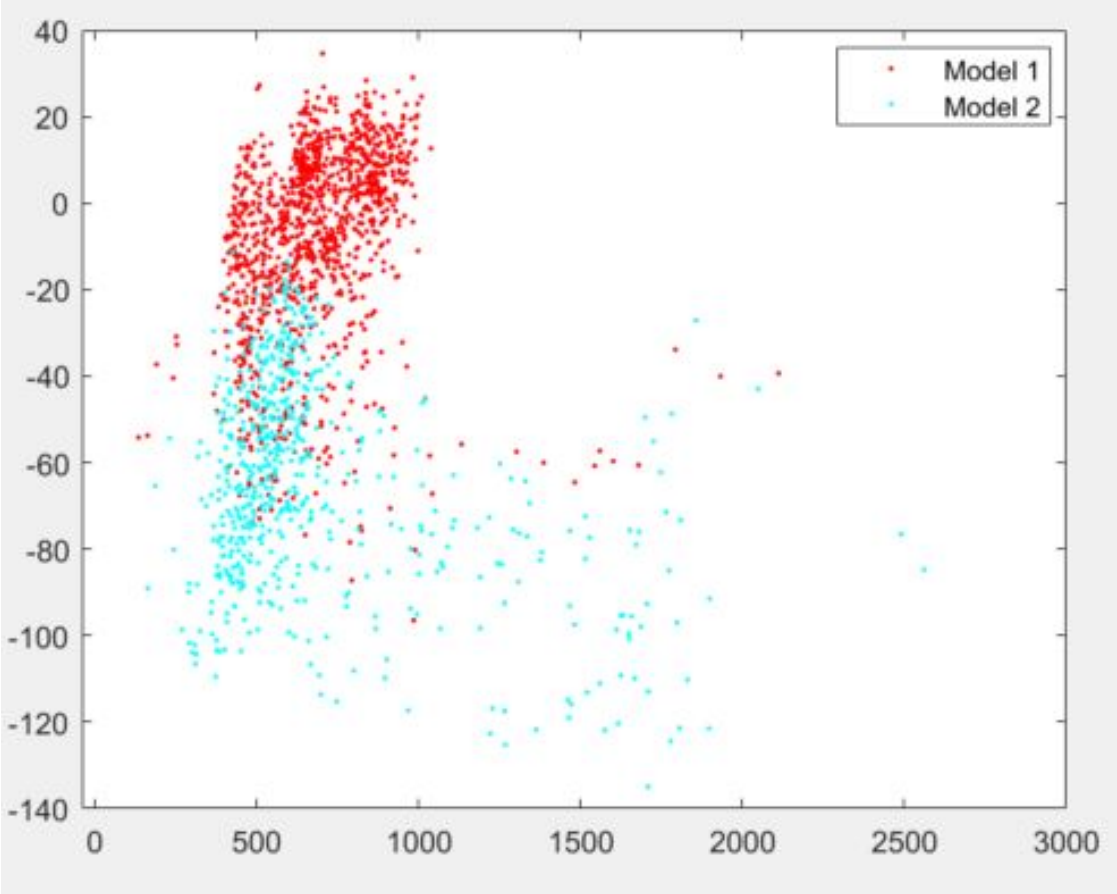

Figure 5-7: Classified points from Isolated Words data in 2D space for the 2-cluster models. Red dots are more likely to be from the GMM corresponding to Vowel and Glide landmarks, and blue dots are more likely to be from the GMM corresponding to Fricatives, Nasals and Stops

From the results in Figure 5-8, it seems that more clusters may model the Fricatives, Nasals, and Stops group more accurately, but does not model the Vowels and Glide landmarks as well as the 1-cluster. This results in a total loss of accuracy, but only by a small amount. The model that classifies Vowel and Glide landmarks took

27 iterations, converged with a log-likelihood of −11226, and took 0.377527 seconds. The second model took 102 iterations, converged with a log-likelihood of −7313.77, and took 0.061313 seconds. This results in a total training time of 0.43884 seconds, and testing took 0.22838 seconds, likely due to the increased complexity of evaluating the larger model. The performance of the 2 cluster models are in Figure 5-8. While it still has very high F1 score, the 2-cluster model does not seem to perform better than the 1-cluster model, possibly because vowels and glides tend to exhibit similar acoustic characteristics, and thus are likely to be clustered around one point and can be represented best by one Gaussian distribution.

#points	VG	FSN
Model 1 (VG)	928	226
Model 2 (FSN)	194	2123

% total	VG	FSN
Model 1 (VG)	26.7358%	6.5111%
Model 2 (FSN)	5.5892%	61.1639%

Accuracy	0.8790
Recall	0.8271
Precision	0.8042
F1	0.8155

Figure 5-8: Results training GMMs with 2 cluster

5.3 Analysis

Below is an analysis of the 169 False Negatives and 228 False Positives that the 1 cluster model for identifying Vowel and Glide landmarks classified incorrectly on the Isolated Words testing data.

Unvoiced Glides are False Negatives

Because formants and amplitude of formants are NaN when the speech is unvoiced, the algorithm will classify all these as Model 2 (Fricatives, Stops, and Nasals), as all vowels and most glides are voiced. However, there is one unvoiced glide, /h/, which ends up always classified incorrectly because of this rule. Out of the 54 False Negatives that result from an unvoiced sound, 28 of these are the 'h' label.

Voiced Vowel and Glide Landmarks

The other False Negatives due to no values for the formant related measurements are Vowel and glide landmarks that are supposed to be voiced but have labels in unvoiced locations. Overall, there are 26 of these errors. These occur when labels are not placed in the middle of the vowel or glide, and instead are placed near the edge where the sound is not as prominent. This means Praat sometimes does not detect F0 and thus the speech signal at that data point is unvoiced. For example, compare the 'l' label in Figure 5-9 (bill), where the glide is successfully detected, with the same label in Figure 5-10 (belt), where it is unsuccessfully detected because the label is placed after the interval where F0 is detected (in blue). For most of the vowel landmarks where this error occurs (12), the labels appear to be placed in the wrong location, as they are being placed at the start of the vowel sound rather than in the middle where the vowel is expressed most prominently. For the glide landmarks (14), the error occurs most commonly when the label is placed too close to the end of the glide sound.

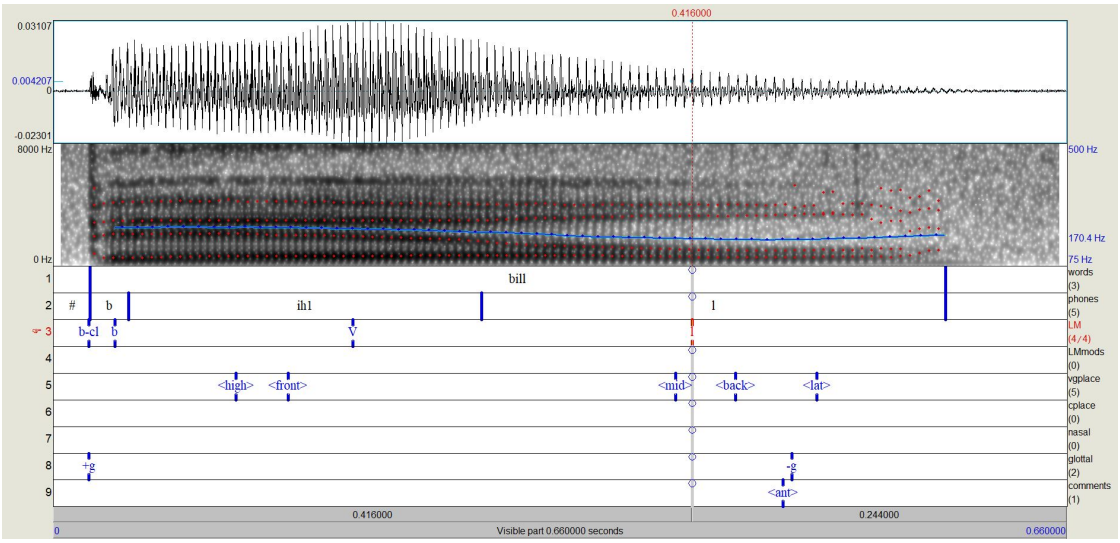

Figure 5-9: Praat Window showing BILL with labels

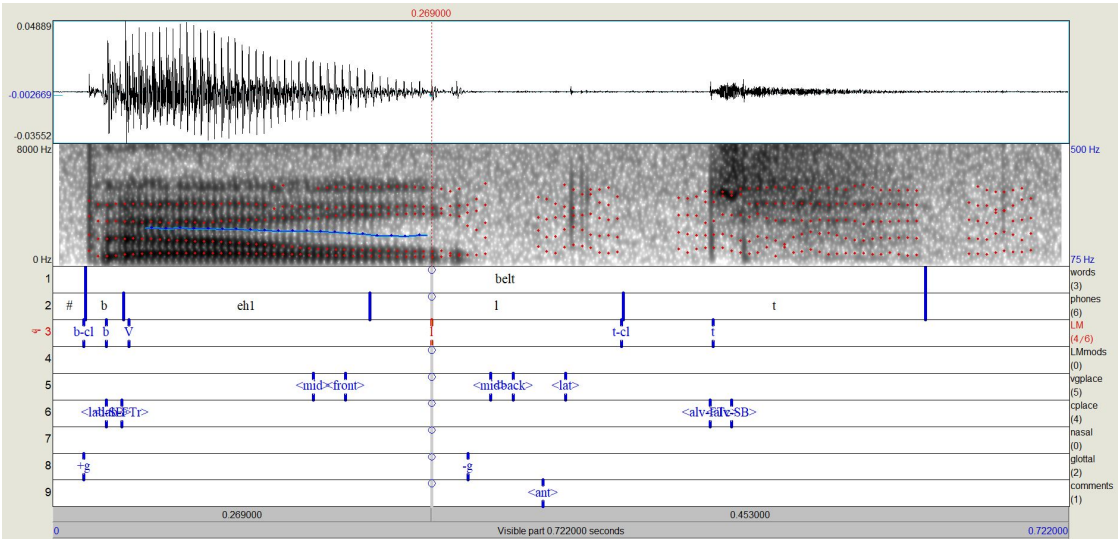

Figure 5-10: Praat Window showing BELT with labels

Other False Negatives

Excluding labelling errors which make up some of the 115 other False Negatives, the most common error in these appears to be incorrectly classifying glides that are faint or short. For example, the 'y' label in Figure 5-11 (yelp) appears to be classified erroneously because Praat fails to detect F1 for a short interval that includes the label. Querying for F1 at that particular location reveals that Praat actually gives the value for F2, as it is the first formant detected by Praat since it did not detect the

actual F1, which is very low. This resulted in the model thinking this data point is very dissimilar to other vowel and glide landmarks, and categorizing it as a consonant landmark.

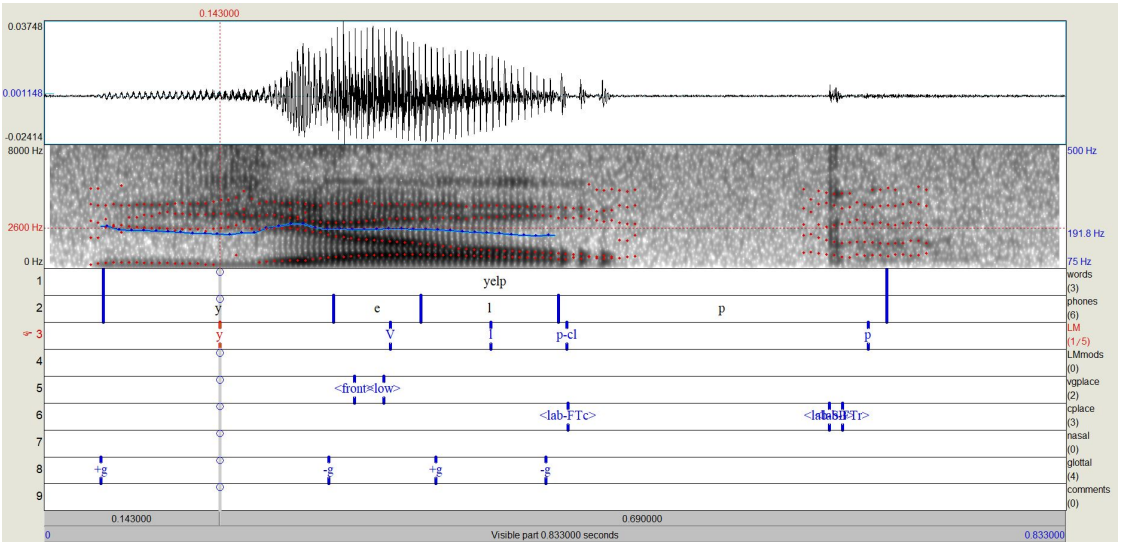

Figure 5-11: Praat Window showing YELP with labels

Fricatives, Stops and Nasals as False Positives

The model resulted in 228 False Positives where Fricatives, Stops, and Nasals were classified as a Vowel or Glide landmark. An in depth analysis of these revealed that they are caused by two main reasons. The first is labelling errors, where the consonant landmark is misplaced on a vowel or a glide, such as the 'n-cl' label in Figure 5-12, which is placed on the end of the preceding vowel. The other is because consonant closures and releases are mostly followed by or preceded by a vowel or glide sound, and by definition mark the edges of the consonant sound, the speech related measurements extracted from the labelled frame often fall towards the vowel/glide side of the boundary. For example, in Figure 5-13, the 'n-cl' label is placed in the correct location but still carries the characteristics of the preceding vowel, such as strong, well defined formants.

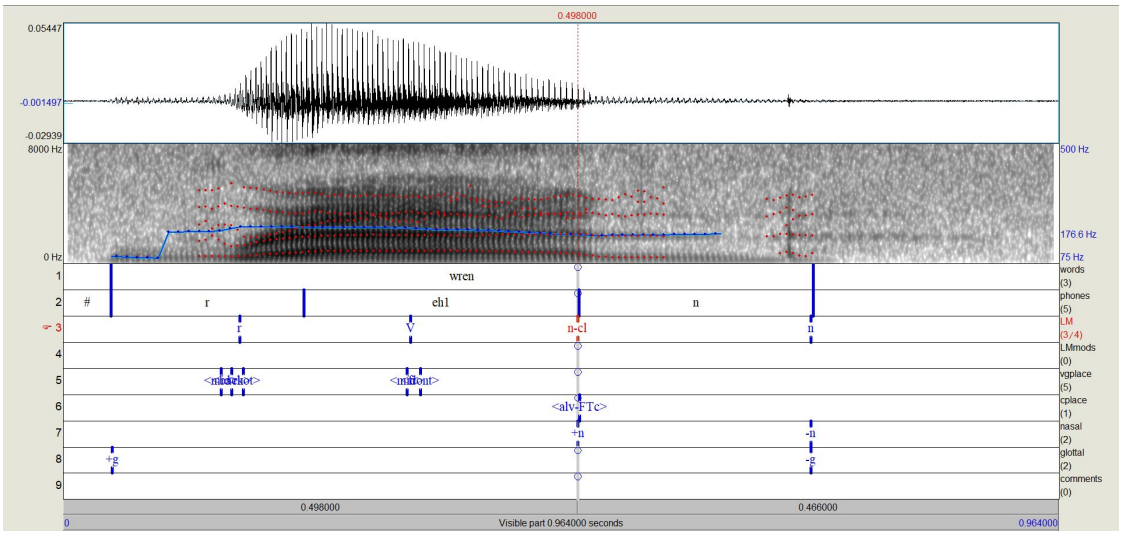

Figure 5-12: Praat Window showing WREN with labels

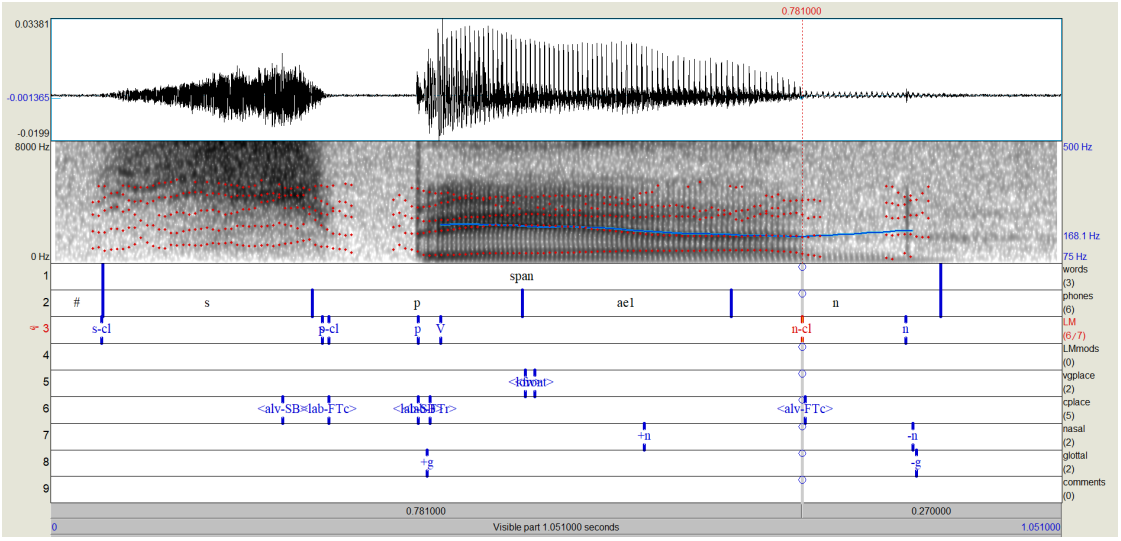

Figure 5-13: Praat Window showing SPAN with labels

5.3.1 Suggestions for Improvement

The analysis of the errors the 1 cluster model made when classifying Vowel and Glide landmarks reveal that perhaps a 1 frame data point is not sufficient for detecting landmarks. This single frame approach has trouble identifying Vowel and Glide landmarks when they are labelled near the beginning or the end of the produced sound, where measurements can be faint or fuzzy. It also finds it difficult to distinguish between a labelled vowel landmark and a labelled consonant landmark that is on the

boundary between a consonant and a vowel. One suggestion for future work is to use a range of values around the label that spans multiple frames. This would capture a range of measurements around the specified frame, and would be slightly more forgiving towards human error in labelling where the landmark is located. Furthermore, it would theoretically be able to distinguish a consonant landmark on the boundary of a vowel or glide landmark, as it would capture the changes in measurement from the start of the range of frames to the end. However, this would increase the number of dimensions in the GMM training space by a large amount, as the total number of dimensions is the number of measurements used multiplied by the number of frames. This may cause training and testing to take more time, and may result in a GMM that fails to converge. One potential solution to this problem would be just to use three frames per landmark location: one for the beginning of the window, one for the location, and one for the end of the window. This would still capture changes in speech related measurements from the start to the end of the range, but would only increase the number of training dimensions by a factor of 3.

Other suggestions for future work include adding more measurement filters. Filters allows the model to be able to focus on different features of the measurements that may not be clear from just the raw measurements. One example of a filter that can be added is one that gets the absolute value of the derivative. Since consonant landmarks generally fall on the boundary between vowels/glides and consonants, the absolute value of the derivative of measurements at that location are likely to be high. Using absolute value captures both upwards and downwards slopes instead of just one direction. Furthermore, Praat can extract the bandwidth of formants, which is a raw measurement that is currently not used but can be added if it is a useful distinguishing factor for certain landmarks.

In the proposed method, the GMMs are trained assuming the covariance matrix is diagonal. This implies that the given measurements, or training dimensions, are uncorrelated. This isn't necessarily true with the provided speech related measurements, as they may have underlying correlations. However, using an unconstrained covariance matrix exponentially increases the number of samples needed for the train-

ing algorithm to converge. Therefore, doing so may achieve better results, but we would need access to much more labelled training data.

Chapter 6

Conclusion

This paper proposed a organized system of steps needed to construct modules for detecting landmark acoustic cues. The system is exemplified here by the development of a module for detecting vowel landmarks. It devised an algorithm for detecting vowel locations in speech files by finding the maximum spectral energy values at formant frequencies within voiced regions of speech, based on observations of speech measurement patterns around labelled vowel landmarks. From this initial module for vowel landmark detection, it conceived a standardized system for building landmark detection modules. First, speech related measurements are extracted using Matlab and Praat, and organized in a data structure by the filename. Then, processing filters can be applied to these raw measurements in order to accentuate certain features. The filters output data in the same overall structure as the raw measurements so that all the measurement data can be processed using the same methods. Next, both the raw and filtered measurements are run against the target landmarks on the ANOVA algorithm which detects correlation between the data points and the labels. A low p-value from the ANOVA algorithm means that the measurement is highly likely to be a good indicator of the presence of the target landmark. Finally, it trains a set of Gaussian Mixture Models using the best measurements confirmed by ANOVA to distinguish between Vowel and Glide landmarks and the other landmarks which are related to consonants. Comparison of the model's outputs to manually labelled test files from the Isolated Words dataset show it to be accurate as well as fast, even

on machines with low resource capabilities. The result is an organized process that can easily be repeated and modified to be able to create an accurate and efficient detection module for any type of landmark cue in speech files. Future work may use these steps to produce Gaussian Mixture Models that separate the Vowel and Glide landmarks, and distinguish the specific consonant landmarks from each other.

Bibliography

[1] Paul Boersma and Vincent Van Heuven. Speak and unspeak with praat. *Glot International*, 5(9/10):341–347, 2001.

[2] Jeung-Yoon Choi. *Detection of consonant voicing: A module for a hierarchical speech recognition system.* PhD thesis, Massachusetts Institute of Technology, 1999.

[3] Edward Flemming. 24.915 / 24.963 linguistic phonetics - the source-filter model of speech production, 2015.

[4] Suryakanth V Gangashetty, Chellu Chandra Sekhar, and Bayya Yegnanarayana. Detection of vowel on set points in continuous speech using autoassociative neural network models. In *Eighth International Conference on Spoken Language Processing*, 2004.

[5] John S Garofolo. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*, 1993.

[6] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[7] Finnian Kelly. *Automatic recognition of ageing speakers.* PhD thesis, Citeseer, 2014.

[8] Sharlene A Liu. Landmark detection for distinctive feature-based speech recognition. *The Journal of the Acoustical Society of America*, 100(5):3417–3430, 1996.

[9] Chi-youn Park. *Consonant landmark detection for speech recognition.* PhD thesis, Massachusetts Institute of Technology, 2008.

[10] Yen-Liang Shue, Patricia Keating, Chad Vicenik, and Kristine Yu. Voicesauce. *p. Program available online at http://www. seas. ucla. edu/spapl/voicesauce/. UCLA*, 2009.

[11] Kenneth N Stevens. Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 111(4):1872–1891, 2002.