An Independent Assessment of Phonetic Distinctive Feature Sets used to Model Pronunciation Variation

Leanne Rolston

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

Masters of Science

University of Washington

2014

Reading Committee:

Gina-Anne Levow, Chair

Richard Wright

Program Authorized to Offer Degree:
Department of Linguistics

University of Washington

Abstract

An Independent Assessment of Phonetic Distinctive Feature Sets used to Model Pronunciation Variation

Leanne Rolston

Chair of the Supervisory Committee:
Gina-Anne Levow
Department of Linguistics

It has been consistently shown that Automatic Speech Recognition (ASR) performance on casual, spontaneous speech is much worse than on carefully planned or read speech by as much as double the word error rate, and that variation in pronunciation is the main reason for this degradation of performance [Ostendorf, 1999]. Thus far, any attempts to mitigate this have fallen well below expectations. Phonetic Distinctive Features show promise from a theoretical standpoint, but have thus far not been fully incorporated into an end-to-end ASR system. Work incorporating distinctive features into ASR is widespread and varied, and each project uses a unique set of features based on the authors' linguistic intuitions, so the results of these experiments cannot be fully and fairly compared. In this work, I attempt to determine which style of distinctive feature set is best suited to model pronunciation variation in ASR based on measures of surface phone prediction accuracy and efficiency of the decision tree model.

Using a non-exhaustive, representative set of phonetic distinctive feature sets, decision trees were trained, one per canonical base form phone, under two experimental conditions: words in isolation, and words in sequence. These models were tested against a comparable held-out test set, and an additional data set of canoni-

cal pronunciations used to simulate formal speech. It was found that a multi-valued articulatory-based feature set provided a far more compact model that yielded comparable accuracy results, while in a comparison of binary feature sets, the model with feature redundancy provided a far more robust model, with slightly higher accuracy and, where it predicted an incorrect phone, it was closer to the actual gold standard phone than the other feature sets' predictions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I'd like to thank my advisor, Gina-Anne Levow, for her patience, mentorship, and encouragement throughout not only this thesis, but my entire core CLMA/CMLS course load, my reader Richard Wright for rekindling my enjoyment of phonetics and phonology, the entire faculty and staff of the UW Linguistics Department for helping me get to this point in my studies. I'd also like to thank the authors whose work I benefited from on this project.

DEDICATION

To my husband Adam, and my sons Hugh and Ian.

Chapter 1

INTRODUCTION

The earliest implementations of Automatic Speech Recognition (ASR) technology recognized specific sounds (Radio Rex, 1911 [Cohen et al., 2004]), isolated words (Bell Laboratories, 1952 [Juang and Rabiner, 2005]), or syllables (RCA Laboratories, 1950's [Juang and Rabiner, 2005]) using the predecessors of our current state of the art ASR technology. As the technology matured, so did our expectations of its capabilities; contemporary ASR systems are expected to discern meaning from spontaneous, casual, and conversational speech with an unconstrained vocabulary. This creates many issues that even state of the art ASR systems are not designed to handle.

It has been consistently shown that ASR performance on casual, spontaneous speech is much worse than on carefully planned or read speech by as much as double the word error rate, and that variation in pronunciation is the main reason for this degradation of performance [Ostendorf, 1999]. There has been a lot of research dedicated to mitigating this problem, but thus far it has failed to yield satisfactory results. It has been speculated that this is due to the fact that acoustic models and lexicons are built using phone-based units, which is not conducive to modeling the types of pronunciation variation found in spontaneous speech [Ostendorf, 1999], which is described in Section 2.1.

The use of phonetic distinctive features has been one attempt to address this issue. Using this method, phones are described in terms of bundles of linguistic features and their values. These features are frequently rooted in articulation, but some do have acoustic correlates. Ideally, the lexicon would also be described in terms of phonetic distinctive features, and mappings from the acoustic signal to the lexicon would be

based on a minimum edit distance measure. A description of such a system is given in [Stevens, 2002]. In theory, this models pronunciation variation accurately especially since it can be used to represent non-canonical forms, that is, forms that do not perfectly match a phone in the phone inventory, which is a major fault with phone based modeling.

There has been a lot of research into the use of phonetic distinctive features in ASR. Feature-based modeling has been proven to improve recognition in a noisy environment [Kirchhoff et al., 2002], and to make more efficient use of training data, allowing us to train more robust models ([Bates, 2003], [Kirchhoff et al., 2002]) among other findings. The issue, however, is that each research project uses a unique set of features chosen to match the authors' intuitions. In this paper, I attempt to find a means to establish which type of feature set (binary or multi-valued, articulatory-based or acoustic, fully specified or underspecified) is best suited to model pronunciation variation caused by spontaneous, casual speech. Using three different variants of distinctive feature sets, and the narrow transcription of the Switchboard Corpus by International Computer Science Institute (ICSI) at the University of Berkeley [Greenberg et al., 1996], I attempt to compare them in terms of surface form prediction accuracy and model efficiency. I am using a decision tree classifier and training models under two experimental conditions: words in isolation and words in sequence. I am then testing those models against a comparable, held-out test set, and an additional simulation comparing each model against a casual and a simulated formal speech style.

While the feature sets' performance does not vary notably in any of the experiments, there were differences in terms of model robustness and compactness. The multi-valued feature set generated the most compact model, and while it was not the most accurate, it was not considerably worse than the best performing feature set. Similarly, a binary model with feature redundancy proved to be the most robust and accurate, but it was the least compact.

It should be noted that the term feature is used in many different contexts in relation to ASR, usually referring to observations in the acoustic signal. In this paper, I am using the term feature in two contexts: phone subunits as part of a distinctive feature inventory, and as units used by a machine learning algorithm. I have made every effort not to use the term ambiguously. Where qualification is needed, I am referring to the phone subunit as a phonetic feature, and the classification unit as a classification feature.

Although these experiments are run on English data, and it is focused on making distinctions in a single language, since it is in the hypothetical context of an ASR system, I am using the more generic term phone rather than phoneme, and following phonetic tradition of enclosing phones in square brackets ([ ]). When the phone is capitalized, it refers to the base form phone. When it is in lower case, it refers to either the gold standard or predicted surface form.

The remainder of this thesis is laid out as follows: section 2 details previous work and more detail on the components I am going to use; section 3 details the data and experimental framework. In section 4, I describe the feature sets I am using, Section 5 shows the analysis of the experimental results. Finally, in section 6 I present my final conclusions and future work on the project. Appendix A has the complete mappings of phones to features for each feature set.

Chapter 2

BACKGROUND

2.1 Pronunciation Variation

There are many causes of variation among speakers. Inter-speaker variation is the result of physiological, regional, socioeconomic, and other factors; basically, regional or social dialects, age, gender and size differences, and other idiosyncrasies that make each speaker unique. Intra-speaker variation refers to differences in speaking styles by a single speaker due to the structure of the language, discourse factors, and speaker emotional state [Wright, 2006].

Inter-speaker variation can be modeled in an ASR acoustic model by training using a variety of speakers of different ages, genders, and regions representative of the population[Wright, 2006]; by using normalization or adaptation techniques to handle differences in vocal tract length and vocal fold size [Benzeghiba et al., 2007]; and by mapping different regional or sociolinguistic variant pronunciations to a single lexical item in the pronunciation dictionary.

Intra-speaker variation is much less systematic and more difficult to model. As the level of formality in speech changes, from more to less formal, the amount of pronunciation variation increases [Greenberg et al., 2003]. In formal speech, the rate of speaking is slowed, articulation is more careful [Strik and Cucchiarini, 1999], and task depending, there may be pauses between words thereby eliminating any cross-word interaction. As speech becomes less formal, the amount of pronunciation variation increases accordingly [Strik and Cucchiarini, 1999].

Types of pronunciation variation include substitution, where a phone becomes another usually through the process of assimilation with a neighbouring phone

[Jurafsky and Martin, 2000]; deletion, where a phone or even a syllable is dropped; and insertion, where a phone is inserted. Table 2.1 shows an example of each of these phenomena.

	Phrase	Canonical Form	Observed Form
Deletion	find him	f ay n d h ih m	f ay n ix m
	draft the	d r ae f t dh iy	d r ae f dh iy
Substitution	set your	s eh t y ow r	s eh ch er
	did you	d ih d y uw	d ih jh ah
Insertion	something	s ah m th ih ng	s ah m p th ih ng

Table 2.1: Examples of Pronunciation Variation from the Switchboard Corpus

Factors that affect the formality of speech include discourse factors, such as how familiar the speakers are with the topic, subject matter, the speaker's emotional or physical state, and even just the fact that the words are spoken in sequence rather than with pauses in between.

To illustrate the difficulties in predicting the variant pronunciations, in their study measuring the effect of speaking rate and word frequency on pronunciation, Fosler-Lussier & Morgan [Fosler-Lussier and Morgan, 1999] found that the rate of phone deletion increased from 9.3% in slow speech to 13.6% in very fast speech. Similarly, the rate of phone substitution increased from 16.9% to 24.2% along the same range of speaking rates. The phone substitutions become less predictable as the speaking rate increased: at the slowest rate of speech, the average number of phone surface realizations per canonical form phone was 3.4, increasing to 4.0 in fast speech [Fosler-Lussier and Morgan, 1999].

These variations are caused by the fact that the articulation of phones requires the synchronization of all the articulators in the vocal tract. When an articulator is not involved in the production of a specific phone, it moves in anticipation of the next phone it will be involved in. This type of asynchronous movement of articulators can

mask the full acoustic consequences of the current phone, though there is evidence that the full articulatory gesture is completed [Livescu, 2005].

2.2 Phonetic Distinctive Feature Theory

Traditionally in phonetics, phones are broadly classified according to their place and manner of articulation, and phonation type, as they are in the International Phonetic Alphabet (IPA). Distinctive features represent a finer grained, abstract representation of phones in that a phone can be described as a bundle of features and each feature's corresponding value. These features and their corresponding values are hypothesized as the mental representation of our phone inventory.

Phonetic Distinctive Feature Theory has a long history, dating back to the works of Trubetzkoy (1939), Jackobson (1949), and Jackobson & Halle (1956) [Moran, 2012]. Initially, the features were rooted in articulation, but have gradually come to incorporate acoustic correlates.

Broad support for Distinctive Feature Theory comes in the fact that many phonological changes act on phones other than those grouped by traditional place or manner of articulation, but rather on phones that share a feature or set of features [Moran, 2012]. An example of this is the devoicing of voiced obstruents (stops and fricatives) at the end of a word, which is attested in many of the world's languages. Further support for the theory lies in the fact that some of the more systematic aspects of disordered speech are also attributable to distinctive feature changes, rather than to random substitutions coincidentally shared by multiple speakers
[Chin and Dinnsen, 1991].

Many different distinctive feature sets have been proposed, some rooted strongly in articulation, others in acoustics or a hybrid of articulatory and acoustic factors [Clements and Hallé, 2010], and some claiming that the underlying mental representation of the phone is not abstract features and their values, but rather articulatory gestures [Livescu, 2005]. Some sets are specific to a particular language, and only aim

to show contrasts within that language, while others strive to be language universal. In a language universal feature set, features that are non-contrastive in a given language form the basis of allophonic variation, such as nasalized vowels in English. The more language specific a distinctive feature set is, the fewer features need to change to change the identity of a phone.

I will further describe the feature sets used in this study in Chapter 4

2.3 The Use of Phonetic Distinctive Features to Model Pronunciation Variation in ASR

Traditional ASR systems regard the phone as the basic unit of speech, and the lexicon as a sequence of phone segments. This is known as the "beads on a string" model [Ostendorf, 1999]. In this model, phones are trained as triphones, a context-dependent model where each phone is trained in the environment of different preceding and following phones which is intended to model the coarticulatory effects of these environments. Because there is a limited amount of training data and an uneven distribution of phones, this may not allow for adequate coverage of the entire phone inventory. This is addressed in some systems by using a factored representation of phones, which is a clustering of phones based on some shared feature values so as to use the limited training data in a more effective way [King et al., 2007]. This method has been found to be able to generalize to unseen contexts effectively [Ostendorf, 2000].

These techniques, while sufficient to handle some forms of variation from the canonical pronunciation, are not really sufficient for modeling pronunciation variation on the scale that we see in casual, spontaneous speech. Since the lexicon is represented as a sequence of phones, the same sequence with a phone substituted or even deleted would not map to that item in the lexicon unless these variant pronunciations were included in the pronunciation dictionary. This is problematic because the more pronunciations included, the greater the chances for confusion with another item in the lexicon [Ostendorf, 2000].

There is widespread agreement that further incorporating distinctive features into ASR would yield benefits ([Ostendorf, 2000], [King et al., 2007], [Stevens, 2002], and many others). One of the arguments supporting their use is that distinctive features are more easily extracted from the acoustic signal than phones: features have more robust and less variant acoustic correlates than phones [Kirchhoff, 2000]; since there are typically fewer features than phones the model for feature extraction should be more robust than for phone extraction [King et al., 2007]; since phones share features training material can be used more efficiently to train more robust models; and in fact, feature recognition rates have been shown to surpass phone recognition rates [Kirchhoff, 2000]. Other arguments supporting the use of distinctive features in ASR reflect the origin of Distinctive Feature Theory and its initial purpose of understanding phonological rules. Distinctive features allow for the modeling of the spreading or overlap of features between adjacent phones, the residual evidence of a deleted phone, and other phenomena that are common in casual spontaneous speech. It is this aspect of modeling pronunciation variation using distinctive features that is the focus of this work.

2.4 Decision Trees and the C4.5 Algorithm

A decision tree is a tree-based predictive model. At each non-terminal node, or decision point, the data instances are divided into subsets based on the value of one of the model's attributes. The attribute used to split the data at each node is chosen based on a maximization of information gain [Witten et al., 2011]. This splitting continues recursively until one of stopping conditions is met:

- All the instances at a node are of the same class.

- There is no additional information gain from splitting these instances further

At this point, the class of the majority of instances is determined, and this class is assigned to the leaf node for that path through the decision tree.

The construction or training of a decision tree is considered greedy in that it will subsume all instances it can at a given node, even at the risk of eliminating the chance of a more accurate classification further down the tree. To avoid over-fitting the training data, an additional stopping condition can be imposed, such as tree depth, number of instances, or some minimal threshold of information gain. Frequently decision trees are further pruned, or cut back so as to be more generalizable.

Once a decision tree model has been built, new instances are *classified* by following a path down the decision tree based on the instance's value of the splitting condition at each node, and the resulting leaf is assigned as its predicted class. This makes decision trees very transparent and easily convertible to rules.

For these experiments, I used Weka's implementation of the C4.5 decision tree classifier, J48 [Hall et al., 2009]. It simplifies the construction of the trees somewhat by ignoring attributes for which the majority of the training instances have the same prediction. Consequently, the decision trees used for these experiments are actually a subtree of the tree in its entirety [Witten et al., 2011].

Chapter 3

METHODOLOGY

3.1 Data

The data used for these experiments was the Switchboard corpus (SWBD), a corpus of conversational telephone speech collected by Texas Instruments in 1992 [Godfrey et al., 1992]. Volunteer subjects, all speakers of American English from various regions and demographics, were randomly paired with other volunteers, given a topic of conversation, and, after a brief opportunity to introduce themselves, were then recorded discussing their given topic for between three and ten minutes. To ensure that the quality of the recordings was not degraded by the telephone network, and that the speakers were recorded on separate channels, the speech signal was recorded directly from the network, bypassing Central Office switches [Godfrey et al., 1992]. The resulting corpus contains about 2500 conversations between 500 speakers. These conversations were then hand transcribed, and time aligned at the word level.

A subset of this corpus was then narrowly phonetically transcribed in 1996 as part of the Switchboard Transcription Project (STP) [Greenberg et al., 1996]. The intention behind this narrow transcription was to investigate how spontaneous casual speech differed from more formal speaking situations. A total of seventy two minutes of the Switchboard corpus, comprising parts of 618 conversations involving 370 speakers representing a wide range of demographics, were transcribed by a team at the International Computer Science Institute (ICSI) at the University of Berkeley under the supervision of Steven Greenberg. These transcriptions captured deviations from the canonical pronunciations, such as phone substitutions, deletions, and insertions.

The transcription system used was a version of ARPABET. It was supplemented

with diacritics to represent phenomena where the resultant surface phone does not exactly match one in the phone inventory. To facilitate mapping these diacritic changes to multiple distinctive feature sets, I have chosen to use a subset of those diacritics. The result of their addition to each distinctive feature set will be described alongside the feature set descriptions in chapter 4. The mapping from ARPABET to the International Phonetic Alphabet (IPA) is given in table 3.1 [1].

ARPABET	IPA	Example	ARPABET	IPA	Example
iy	i	s**ee**n	ih	ɪ	s**i**n
ey	e	h**ay**	eh	ɛ	r**e**d
ae	æ	b**a**t	aa	ɑ	dr**o**p
ao	ɔ	c**o**ffee	ow	o	m**oa**n
ah	ʌ	m**u**d	uw	u	**oo**ze
uh	ʊ	w**oo**d	ux	ʉ	y**ou***
ax	ə	b**a**nan**a**	er	ɚ	nutt**er**
ix	ɨ	p**o**tato*	aw	ɑw	c**ou**ch
ay	ɑj	**I**	oy	ɔy	t**oy**
hh	h	**h**am	w	w	**w**eave
y	j	**y**ell	r	ɹ	**r**un
dx	ɾ	bu**tt**er*	q	ʔ	bu**tt**on
nx	ɾ^{n}	ya **kn**ow*	hw	ʍ	**wh**ether
lg	ɬ	resu**l**t			
l	l	**l**ift	el	l̩	pood**le**
m	m	lla**m**a	em	m̩	the**m***
n	n	**n**ose	en	n̩	garde**n**
ng	ŋ	si**ng**	eng	ŋ̩	buyi**ng***
v	v	**v**ine	f	f̩	**f**ine
dh	ð	ei**th**er	th	θ	e**th**er
z	z	**z**oo	s	s	**s**ue
zh	ʒ	a**z**ure	sh	ʃ	**sh**ut
b	b	**b**ind	p	p	**p**ined
d	d	**d**aisy	t	t	**t**azer
g	g	**g**irl	k	k	**k**ite
jh	dʒ	**g**in	ch	tʃ	**ch**in

Table 3.1: Switchboard ARPABET to IPA Mappings

Note that the diphthongs [aw], [ay], and [oy] are split into two segments so as to more accurately map distinctive feature values to those phones.

[1] mapping from [Bates, 2003]

Since the STP phone-to-word mappings are stored in two separate files, one for the words and their end timestamps, and one for the phones and their timestamps, phones had to be aligned with the corresponding word. The alignments were then manually edited to fix instances where a non-speech 'word', designated as 'H#', was not aligned with a non-speech sound, designated as 'h#'. Where this caused an obvious misalignment of a phone to its corresponding word in timestamps adjacent to the H# was also manually fixed. No other misalignments were fixed.

3.2 Experimental Framework

Three phonetic feature sets were chosen to model pronunciation variation on a single data set. Both the Hayes Distinctive Feature Set, further described in section 4.2, and the Stevens set, further described in section 4.1 are binary distinctive feature sets descendant from Chomsky and Halle's *The Sound Pattern of English* (1968). They differ in the actual distinctive features chosen, as well as the hierarchical structure and how they represent dependencies between different features. In the implementations of the sets that I have chosen to use, they also differ in how they regard features that are not applicable to a specific phone. These are either underspecified in that their status does not change the identity of the phone within the specific language (i.e. the feature is not contrastive), or the articulator relating to the feature is not involved in the production of the phone. The Stevens Distinctive Feature Set regards these states as being distinct, while the Hayes set does not distinguish between them. The third feature set, further described in section 4.3, is an implementation of Browman and Goldstein's Articulatory Phonology by Livescu [Livescu, 2005]. It is more accurately considered a gestural feature set in that all of the features relate to specific positions of the articulators, while the Hayes and Stevens sets are phonetic feature sets.

Using each distinctive feature set, decision trees were trained, one per canonical base form phone, with the form of the phone in the observed surface form as input, and the surface form predictions as output. Classification features were based on Bates

(2003) [Bates, 2003] and will be described more fully in section 3.3. Canonical pronunciations and syllabification are based on a version of the Carnegie Mellon University (CMU) Pronouncing Dictionary, version 0.6 [Bartlett et al., 2009b], that was augmented with syllabification by Bartlett, Kondrak, and Cherry [Bartlett et al., 2009a], who used a language independent, Support Vector Machine Hidden Markov Model (SVM-HMM) to tag phonemes according to their syllabic role (onset, nucleus, coda), based on a simplified sonority scale supplemented with some language specific constraints [Bartlett et al., 2009a]. The version of the STP data set used is the 1997 workshop format. I used the given test/train split for that data set, with parts of 610 conversations in the training set, and seven conversations in the test set.

Figure 3.1 shows the work flow for all experiments. The mapping of phones to sets of features is included in Appendix A. The phone alignments are based on the minimum edit distance dynamic programming algorithm as presented in Jurafsky & Martin (2000) [Jurafsky and Martin, 2000].

Two models were trained, one using words as a single unit and ignoring inter-word pronunciation variation, and one using speaking turns as a unit which acknowledges the effect of adjacent words on pronunciation. Where the model was considered non-contrastively, I used the latter set since it had more contextual data and represented a more realistic ASR scenario. The first model was used contrastively to compare each feature set's modeling of canonical pronunciation against observed surface forms. The motivation behind this comparison was to assess whether any of the feature sets was particularly suited to model the range of pronunciation variation that occurs in spontaneous speech, or if they were all better suited for more systematic forms of pronunciation variation, such as is found in more formal speech.

3.3 Classification Feature Selection

As previously stated, classification features were based on Bates (2003) [Bates, 2003]. Since the focus of this work is to determine which type of distinctive feature set is

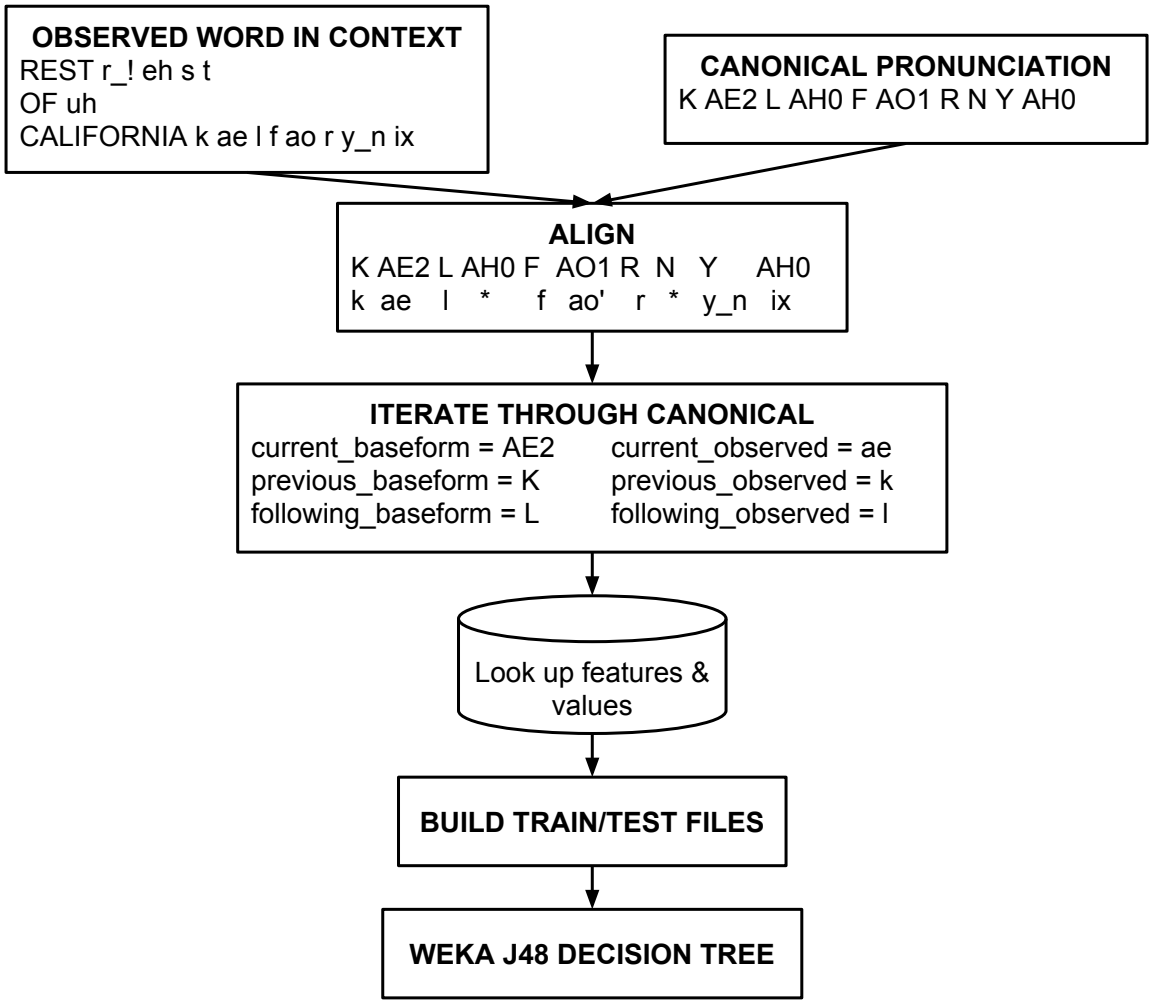

Figure 3.1: Experimental Work Flow

computationally best suited for the task of modeling and predicting pronunciation variation in ASR, I have chosen to focus mainly on the classification features relating to the phonetic environment rather than higher linguistic units such as syntax and discourse, or other factors such as speaking rate or word predictability. These factors may be included in future related work.

Table 3.2 describes the features and their possible values.

Feature	Values
Current canonical phone features	see section 4
Previous canonical phone features	see section 4
Following canonical phone features	see section 4
Current observed form phone features	see section 4
Previous observed form phone features	see section 4
Following observed form phone features	see section 4
phone location in word	beginning, end, medial
phone location in syllable	onset, nucleus, coda
syllable lexical stress	primary, secondary, none

Table 3.2: Surface Level Phone Prediction Features

To avoid experimenter bias regarding which phonetic features may influence adjacent phones, I am including the entire feature matrix for a given phone. I include both the features of the canonical form and the observed form since both the intended target articulation and the actual articulation could have an affect on adjacent phones. Since the data set I am using includes diacritic information, the influence of these is incorporated into the feature values of the observed feature forms. Out of simplicity, I am restricting myself to a three-phone window of current, previous, and following phones, ignoring any long distance influence phones may have on each other.

The phone location within a syllable is important for several reasons. First, there is the finding that, overall, onset consonants are pronounced closer to their canonical form than other consonants [Greenberg, 1999]; this tendency increases in the absence of a canonical coda [Greenberg, 1999], and in syllables with full stress

[Greenberg et al., 2003]. Secondly, the type of deviation from the canonical form tends to depend on the position within the syllable [Greenberg et al., 2003]: substitutions occur most often in the nucleus while deletions in the nucleus are very rare [Greenberg et al., 2003]; insertions, while rare overall, are concentrated in the syllable onset [Greenberg et al., 2003]; deletions, and consonantal substitutions due to assimilation happen most often in the syllable coda. Thirdly, the likelihood of a canonical pronunciation permeates throughout a syllable; if an onset is pronounced close to canonical form, this increases the likelihood that the nucleus and the coda will also be [Greenberg, 1999].

Lexical stress has also been found to have an effect on pronunciation variation. A greater degree of lexical stress on a syllable increases the likelihood that all phones in that syllable will be pronounced closely to canonical form; a lesser degree of stress increases the chance of substitutions in the nucleus and deletions in the coda [Greenberg et al., 2003]. Lexical stress also has has an influence on the articulatory space of some phones, for example, there is a tendency for vowel raising in unaccented syllables [Greenberg et al., 2003].

3.4 Assessment

The three feature sets are compared to determine their suitability to the task of modeling pronunciation variation using measures of surface phone prediction accuracy, characteristics of the decision tree models to estimate model efficiency, rates of phone class confusion, and a feature edit distance measure, described in section 5.4. These models are tested against a held out test set.

Chapter 4

CANDIDATE DISTINCTIVE FEATURE SETS

The feature sets I have chosen represent various flavours of Distinctive Feature Theory. Stevens, further described in section 4.1, is a binary feature geometry, which although articulatory in nature, is also firmly rooted in acoustics. Hayes is also a binary distinctive feature set, and while it is not explicitly represented as a hierarchy, does have feature dependencies which are described in section 4.2. The Livescu Feature Set, described in section 4.3, is based on Browman and Goldstein's Articulatory Phonology [Livescu, 2005] in which the gesture, rather than an underlying target set of phonetic features, is seen as the underlying basis for speech. For this reason, it is more commonly referred to as a gestural feature Set.

4.1 Stevens

The Stevens Distinctive Feature Set is designed for use in Automatic Speech Recognition (ASR) [Stevens, 2002]. In presenting the feature set with very strong acoustic justification for each feature, Stevens presents a use case in which it is used in a Landmark or Event based ASR system, which requires two passes over the speech signal. The first pass detects frequency peaks, valleys, and discontinuities in the speech signal, which represent acoustic landmarks and can be mapped to articulator-free features (see section 4.1.1), and the second pass detects clues in the vicinity of those landmarks to estimate articulator-bound features (see section 4.1.2). From these, bundles of distinctive features are generated, and mapped to a lexicon with the same representation. While it is not specifically limited to English, it does lack the features to mark contrasts in many of the world's languages.

The implementation of the Stevens Distinctive Feature Set that I used is from Bates (2003) [Bates, 2003] with some modifications. It includes the following feature values:

- **+**: the feature is *on* for the phone
- **-**: the feature is *off* for the phone
- **0**: the feature's value does not change the identity of the phone (underspecified)
- **x**: the feature's value is not applicable to the phone

4.1.1 Articulator-Free Features

Articulator-free features do not refer to a specific articulator, but rather to the degree of vocal tract constriction and the acoustic consequences of such constrictions [Stevens, 2002]. In broadest terms, the top level of division divides the phone set into vowels and consonants. A third category of phone can also be determined at this level; glides are the result of a constriction in the vocal tract that is not sufficient to create an acoustic discontinuity.

Feature	Vowels	Glides	Consonants
vocalic	+	-	-
consonantal	-	-	+

Table 4.1: Stevens' High Level Divisions of Segments

Vowels are recognizable in the acoustic signal by a maximum in the low to mid frequency spectrum amplitude. True consonants, on the other hand, are identified by a lowering of that spectrum amplitude, leading to one discontinuity caused by the vocal tract constriction, and another discontinuity when the constriction is released. Glides show the same lowering of the spectrum amplitude without the discontinuity.

There are other articulator-free features to further describe [+consonantal] features that relate to the discontinuity. The feature [continuant] refers to the completeness of the vocal tract closure: [+continuant] refers to an incomplete closure that generates continuous turbulence, while [-continuant] refers to a complete closure and its related absence of signal. [-continuant] segments are subdivided into [+sonorant], which have no pressure buildup behind the closure, and [-sonorant] which do have pressure buildup. [+continuant] segments which are produced using the blade of the tongue, are divided into [+strident] reflecting a tongue position that directs the airflow against the lower teeth and the resulting high frequency sound hissing noise, and [-strident], which do not. This hierarchy is presented in figure 4.1.

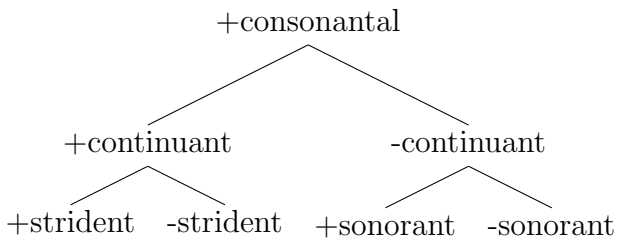

Figure 4.1: Stevens' Subdivision of +consonantal Segments

4.1.2 Articulator-Bound Features

Articulator-bound features refer to the state or action of a specific articulator. This feature set considers seven articulators: lips, tongue blade, tongue body, soft palate, pharynx, glottis, and vocal folds, each with specific features associated with them. These articulators and their features are shown in table 4.2.

4.1.3 Feature Matrices

Vowels are defined using the features associated with the tongue body, lips, and pharynx, as listed in table 4.2. Monopthongs are defined as a single segment, while

Articulator	Associated Feature
lips	[round]
tongue blade	[anterior]
	[distributed]
	[lateral]
tongue body	[high]
	[low]
	[back]
soft palate	[nasal]
pharynx	[advanced tongue root]
glottis	[spread glottis]
	[constricted glottis]
vocal folds	[stiff vocal folds]

Table 4.2: Stevens' Articulators and their defining articulator-bound features

diphthongs are defined as two adjacent segments, the second segment being a glide.

Table 4.3 shows an abbreviated feature matrix for the monopthong vowels. In addition to these features, they share the following features in common: [+vocalic], [-consonantal], [+continuant], [+sonorant], [xstrident], and [+delayed release]. The complete feature matrix can be found in table A.1 in Appendix A.2.

Glides are defined using the features associated with the lips, tongue blade, tongue body, soft palate, pharynx, and glottis. Table A.2 in Appendix A.2 shows the features matrix for the diphthongs and glides.

In addition to the articulator-free features specified in figure 4.1, consonants are further specified in terms of the features associated with the primary articulator that is used to form the constriction, and, if relevant, a secondary articulator. Only three articulators can form these constrictions: the lips, the tongue blade, and the tongue body, while there are two secondary articulators that may be engaged to form contrastive segments in English: the glottis and the vocal folds. Each of these has an associated set of defining features, some of which must be defined for a given articulator, while other features may remain unspecified. When features are unspecified,

ARPABET symbol	iy	ih	ey	eh	ae	aa	ao	ow	ah	uw	ux	uh	ax	ix	er
lips	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0
tongue blade	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
tongue body	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
round	-	-	-	-	-	-	+	+	-	+	+	+	0	0	0
anterior	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
distributed	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lateral	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
high	+	+	-	-	-	-	-	-	-	+	+	+	-	0	0
low	-	-	-	-	+	+	+	-	-	-	-	-	-	0	0
back	-	-	-	-	-	+	+	+	+	+	-	+	0	-	0
nasal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
advanced tongue root	+	-	+	-	-	-	-	+	-	+	+	-	-	-	-
constricted tongue root	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-
spread glottis	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
constricted glottis	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
stiff vocal folds	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 4.3: Abbreviated Stevens Distinctive Features for Vowels

it could be for one of two reasons: 1) the articulation is unlikely or impossible (i.e. the articulator is moving toward another target), 2) the feature is not contrastive in English. In our nomenclature, the unlikely articulations are indicated by an **x**, while the underspecified, non-contrastive features are indicated by a **0**.

All consonants share the following features in common: [-vocalic], [+consonantal], [xadvanced tongue root], and [xconstricted tongue root].

Table 4.4 shows the abbreviated feature matrix for consonants that do not have a corresponding syllabic variant. The full feature matrix is in table A.4 in Appendix A.2. The feature matrix for the consonants that have a corresponding syllabic variant are given in table A.3, also in Appendix A.2.

ARPABET symbol	v	dh	z	zh	f	th	s	sh	b	d	g	p	t	k	jh	ch
continuant	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-
sonorant	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
strident	+	-	+	+	+	-	+	+	x	x	x	x	x	x	+	+
delayed release	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-
lips	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
tongue blade	x	+	+	+	x	+	+	+	x	+	x	x	+	x	+	+
tongue body	x	x	x	x	x	x	x	x	x	x	+	x	x	+	x	x
round	-	-	-	-	-	-	-	-	-	+	-	-	+	-	-	-
anterior	x	+	+	-	x	+	+	-	x	+	x	x	+	x	-	-
distributed	x	+	-	+	x	+	-	+	x	-	x	x	-	x	-	-
lateral	x	-	-	-	x	-	-	-	x	-	x	x	-	x	-	-
high	x	x	x	x	x	x	x	x	x	x	+	x	x	+	x	x
low	x	x	x	x	x	x	x	x	x	x	-	x	x	-	x	x
back	x	x	x	x	x	x	x	x	x	x	+	x	x	+	x	x
nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
spread glottis	-	-	-	-	+	+	+	+	-	-	-	+	+	+	-	+
constricted glottis	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
stiff vocal folds	-	-	-	-	+	+	+	+	-	-	-	+	+	+	-	+

Table 4.4: Stevens Distinctive Features for Consonants Without Corresponding Syllabic Consonants

4.1.4 Diacritics

Table 4.5 details the feature changes caused by the diacritics. This is just a subset of the total set of diacritics used in the Switchboard Transcription Project (STP), and were chosen because of their ease of portability into different feature sets.

Diacritic	Meaning	Feature Changes
n	nasalization of phone	*nasal* to +
vd	voicing of voiceless phone	*stiff vocal folds* to -
vl	devoicing of voiced phone	*stiff vocal folds* to +

Table 4.5: Stevens Diacritic Effects

4.2 Hayes

Hayes presents a much more general use distinctive feature set. Like its predecessor, Chomsky and Halle's *The Sound Pattern of English* (1968), it is strongly rooted in articulation, but does factor in acoustics where necessary. While not explicitly represented as a feature geometry, there are some inherent feature dependencies among the place features, further described in section 4.2.2.

The complete mapping of features to phones came from Hayes' web site [1], where I used only the subset of the phones that overlaps with the phones used in the Switchboard Transcription Project (STP) data, as given in table 3.1.

The Hayes Feature Set has the following feature values:

- **+**: the feature is *on* for that phone
- **-**: the feature is *off* for that phone
- **0**: means the feature is not relevant to the phone in question, or the phone is said "not to care" about its value [Hayes, 2011].

Features are broadly categorized into manner features, laryngeal features and place features.

4.2.1 Manner Features

The features [sonorant], [approximant], [consonantal], and [syllabic] are based on the sonority hierarchy, which ranks the traditional manners of articulation based on acoustic loudness, or sonority, and is in table 4.6. Sonority decreases from left to right.

---

[1] http://www.linguistics.ucla.edu/people/hayes/IP/Files/FeaturesDoulosSIL.xls, retrieved 11-10-13

Vowels	**Glides**	**Liquids**	**Nasals**	**Obstruents**
[+syllabic]	[-syllabic]			
[-consonantal]		[+consonantal]		
[+approximant]			[-approximant]	
[+sonorant]				[-sonorant]

Table 4.6: Hayes' Manner Features in the Sonority Hierarchy

The feature [sonorant] refers to the degree of constriction in the vocal tract, and the corresponding amount of acoustic energy generated by that constriction. [+sonorant] segments have greater acoustic energy than [-sonorant], while [-sonorant] segments have a closure narrow enough to allow for an buildup of air pressure in the vocal tract. Stops, fricatives, and affricates (obstruents) are [-sonorant], while all other segments are [+sonorant].

The feature [syllabic] defines a phone's suitability to serve as the nucleus of a syllable. According to the sonority hierarchy, the phone with the highest sonority serves as the syllable nucleus. Sonority decreases as you move away from the nucleus in both directions. Those segments that can serve as a syllable nucleus are [+syllabic], and include vowels, syllabic liquids, and syllabic nasals in English. All other segments are [-syllabic].

[consonantal] refers to whether a segment has an audible constriction in the vocal tract: liquids, nasals, and obstruent segments are [+consonantal], while vowels and glides are [-consonantal].

[approximant] marks the cutoff point between vowels, glides, and liquids, which are [+approximant], and nasals and obstruents, which are [-approximant]. In terms of articulation, it refers to the approach of an articulator toward its target that is not sufficient to generate turbulent airflow.

To further distinguish stops, fricatives, and affricates, there are two more manner features, [continuant] and [delayed release]. The feature [continuant] refers to

the degree of closure of the oral portion of the vocal tract. [+continuant] refers to segments that do not have a complete closure: fricatives, liquids, glides, and vowels; [-continuant] segments: stops, affricates, and nasals, do. The delay in [delayed release] refers to the period of semi-closure where frication occurs. Stops, that lack any frication are [-delayed release] while affricates and fricatives are [+delayed release].

Finally, to distinguish the various types of [r], the manner features [trill] and [tap] are included. Taps are generated at the alveolar place of articulation, and are paired with the place feature [+anterior], while flaps are retroflex, and are paired with [-anterior].

Table 4.7 shows the manner features for a cross section of phones. The complete feature charts can be found in A.2 in Appendix A.2.

ARPABET symbol	iy	ey	uw	hh	w	l	el	v	f	b	p
consonantal	-	-	-	-	-	+	+	+	+	+	+
syllabic	+	+	+	-	-	-	+	-	-	-	-
sonorant	+	+	+	-	+	+	+	-	-	-	-
continuant	+	+	+	+	+	+	+	+	+	-	-
delayed release	0	0	0	+	0	0	0	+	+	-	-
approximant	+	+	+	-	+	+	+	-	-	-	-
tap	-	-	-	-	-	-	-	-	-	-	-
trill	-	-	-	-	-	-	-	-	-	-	-
nasal	-	-	-	-	-	-	-	-	-	-	-

Table 4.7: Hayes Manner Features

4.2.2 Place Features

The place features are determined by the primary articulator, or articulators, used in generating a phone, and that articulator's location or type of movement within the vocal tract.

[labial] refers to articulation involving the lips. [+labial] phones are further described depending on the type of lip movement required to produce a sound. Lip rounding is represented by [+labial] [+round] while contact between the lips and

the teeth is represented by [+labial] [+labiodental]. All other gestures are marked [-labial].

[coronal] refers to whether phones are produced using the tongue blade or tip ([+coronal]), or not ([-coronal]). [+coronal] phones are further defined with the features [+anterior] if the tongue makes contact forward of the alveolar ridge and [-anterior] if the contact is farther back; [+distributed] if it is the tongue blade making contact with the roof of the mouth and [-distributed] if it is the tongue blade; [+strident] if the tongue directs the airflow through a channel and against the back of the teeth and [-strident] if it does not; and [+lateral] if the airflow through the oral cavity is directed around the side of the mouth and [-lateral] if it is not. .

The feature [dorsal] marks whether a phone is produced using the tongue body. [+dorsal] segments are further described with the features [high], [low], [front], [back], and [tense]. [high] and [low] make a three-way distinction in height, while [front] and [back] allow for a three-way horizontal distinction. That there is a five-way distinction in height in most vowel systems is addressed with the feature [tense], which refers to the position of the tongue root, [+tense] being slightly forward of [-tense] segments.

Table 4.8 shows the place features for a cross section of phones. The complete feature charts can be found in Appendix A.2

4.2.3 Laryngeal Features

The laryngeal features are [voice], [spread glottis] and [constricted glottis].

[+voice] has both an articulatory and acoustic correlate. Articulatorily, it refers to whether the vocal folds vibrate regularly. Acoustically, it refers to the corresponding periodicity in the signal that results from that vocal fold vibration. [-voice] refers to the absence of that vocal fold vibration.

[+spread glottis] refers to a state where the vocal folds are held far apart. This feature is present in the segment [h], for breathy vowels, and aspirated consonants.

ARPABET symbol	iy	ey	uw	hh	w	l	el	v	f	b	p
labial	-	-	+	-	+	-	-	+	+	+	+
round	-	-	+	-	+	-	-	-	-	-	-
labiodental	-	-	-	-	-	-	-	+	+	-	-
coronal	-	-	-	-	-	+	+	-	-	-	-
anterior	0	0	0	0	0	+	+	0	0	0	0
distributed	0	0	0	0	0	-	-	0	0	0	0
strident	0	0	0	0	0	-	-	0	0	0	0
lateral	-	-	-	-	-	+	+	-	-	-	-
dorsal	+	+	+	-	+	-	-	0	0	-	-
high	+	-	+	0	+	0	0	0	0	0	0
low	-	-	-	0	-	0	0	0	0	0	0
front	+	+	-	0	-	0	0	0	0	0	0
back	-	-	+	0	-	0	0	0	0	0	0
tense	+	+	+	0	+	0	0	0	0	0	0

Table 4.8: Hayes Place Features

[+constricted glottis], on the other hand refers to a state where the vocal folds are held tightly together, allowing little or no airflow through. This is the case for glottal stops and other glottalized sounds.

Table 4.9 shows the place features for a cross section of phones. The complete feature charts can be found in Appendix A.2

ARPABET symbol	iy	ey	uw	hh	w	l	el	v	f	b	p
voice	+	+	+	-	+	+	+	+	-	+	-
spread glottis	-	-	-	+	-	-	-	-	-	-	-
constricted glottis	-	-	-	-	-	-	-	-	-	-	-

Table 4.9: Hayes Laryngeal Features

4.2.4 Diacritics

Table 4.10 details the feature changes caused by the diacritics. This is just a subset of the total set of diacritics used in the Switchboard Transcription Project (STP), and were chosen because of their ease of portability into different feature sets.

Diacritic	Meaning	Feature Changes
n	nasalization of phone	*nasal* to +
vd	voicing of voiceless phone	*voice* to +
vl	devoicing of voiced phone	*voice* to -

Table 4.10: Hayes Diacritic Effects

4.3 Livescu

The Livescu Feature Set, based on the vocal tract variables from Browman and Goldstein's Articulatory Phonology [Livescu, 2005] sees the articulatory gesture as the underlying phonological unit, and utterances as a series of coordinated, potentially overlapping gestures. These gestural-based features refer to the location and degree of constriction of the major articulators in the vocal tract and are shown in table 4.11.

The features are grouped by articulator: lips, tongue tip, tongue blade, velum, and glottis.

The lip features relate to the position and the degree of opening of the lips. The feature LIP-LOC refers to the horizontal displacement of the lips, from rounded (PRO), neutral (LAB), to the position the lips are in for labio-dental phones (DEN). LIP-OPEN relates to how far apart the lips are from each other, and the values range from closed (CL), critical (CR), narrow (NA), and wide (WI).

The tongue tip features refer to the location of contact of the tongue tip and the amount of constriction generated caused by this contact. The TT-LOC values are inter-dental (DEN), alveolar (ALV), palato-alveolar (P-A), and retroflex (RET). The TT-OPEN values range from a complete constriction (CL), a critical constriction such as one that creates frication (CR), narrow (NA), medium-narrow (M-N), medium (MID), and wide (WI).

Similarly, the tongue blade features refer to the location of contact or near contact of the tongue blade and the amount of constriction generated by that contact. The values for TB-LOC are palatal (PAL), velar (VEL), uvular (UVU), which is the

Feature	Description	Value
LIP-LOC	position of the lips	PRO = protruded (rounded) LAB = labial (default/neutral position) DEN = dental (labio-dental position)
LIP-OPEN	degree of opening of the lips	CL = closed CR = critical NA = narrow WI = wide
TT-LOC	location of the tongue tip	DEN = interdental ALV = alveolar P-A = palato-alveolar RET = retroflex
TT-OPEN	degree of opening of the tongue tip	CL = closed CR = critical NA = narrow M-N = medium-narrow MID = medium WI = wide
TB-LOC	location of the tongue body	PAL = palatal VEL = velar UVU = uvular (default/neutral position) PHA = pharyngeal
TB-OPEN	degree of opening of the tongue body	CL = closed CR = critical NA = narrow M-N = medium narrow MID = medium WI = wide
VEL	state of the velum	CL = closed (non-nasal) OP = open (nasal)
GLOT	state of the glottis	CL = closed (glottal stop) CR = critical (voiced) OP = open (voiceless)

Table 4.11: Articulatory Phonology Features and Values

neutral position, and pharyngeal (PHA). TB-OPEN values are closed (CL), as for a stop consonant, critical (CR), or enough to create frication, narrow (NA), such as for [y], medium-narrow (M-N), medium (MID), and wide (WI).

The velar (VEL) features refer to the state of the velum. Closed (CL) refers to a non-nasal phone while open (OP) refers to a nasal.

Finally, the glottal (GLOT) values refer to the state of the glottis. Its values are closed (CL) as for a glottal stop, critical (CR) as for a voiced phone, and open (OP), as for a voiceless phone.

The feature to phone mappings were compiled by Livescu's team based on the literature of Browman and Goldstein as well as some general phonological literature and X-ray tracings of speech articulation [Livescu, 2005]. Table 4.12 shows the feature mappings for a subset of the phones. The complete mapping of phones to features is given in Appendix A.3.

ARPABET	iy	uw	hh	w	v	f	b	p
LIP-LOC	LAB	PRO	LAB	PRO	DEN	DEN	LAB	LAB
LIP-OPEN	WI	NA	WI	NA	CR	CR	CR	CR
TT-LOC	ALV	P-A	ALV	P-A	ALV	ALV	ALV	ALV
TT-OPEN	M-N	WI	MID	WI	MID	MID	MID	MID
TB-LOC	PAL	VEL	UVU	UVU	VEL	VEL	UVU	UVU
TB-OPEN	M-N	NA	MID	NA	MID	MID	WI	WI
VEL	CL	CL	CL	CL	CL	CL	CL	CL
GLOT	CR	CR	WI	CR	CR	WI	CR	WI

Table 4.12: Livescu Abbreviated Phone to Feature Mappings

4.3.1 Diacritics

Table 4.13 details the feature changes caused by the diacritics. This is just a subset of the total set of diacritics used in the Switchboard Transcription Project (STP), and were chosen because of their ease of portability into different feature sets.

Diacritic	Meaning	Feature Changes
n	nasalization of phone	*VEL* to OP
vd	voicing of voiceless phone	*GLOT* to CR
vl	devoicing of voiced phone	*GLOT* to OP

Table 4.13: Articulatory Phonology Diacritic Effects

Chapter 5

ANALYSIS

5.1 Data Analysis

As stated in section 3.1, the experiments were run on the International Computer Science Institute (ICSI) narrowly transcribed subset of the Switchboard Transcription Project (STP) data [Greenberg et al., 1996], and specifically the 1997 Workshop format[1]. I used the given train/test split, with the training set having 736 speaker participation turns over 610 conversations (some speakers were involved in more than one conversation), and the test set having 12 speaker participation turns over 6 conversations.

Table 5.1 shows the count of each phone in both the training set and the test set. The symbol '?' represents an unknown sound and 'h#' represents a non-speech sound.

Phone	Training Set	Test Set	Phone	Training Set	Test Set
?	42	34	iy	754	517
aa	328	281	jh	94	67
ae	606	284	k	557	479
ah	586	417	l	461	340
ao	261	168	lg	115	59
aw	131	90	m	493	399
ax	833	644	n	1206	757

Table 5.1 Continued on next page

[1]http://www1.icsi.berkeley.edu/Speech/stp/, retrieved 17 December, 2013

Table 5.1 – continued from previous page

Phone	Training Set	Test Set	Phone	Training Set	Test Set
ay	484	344	ng	202	137
b	324	249	nx	128	113
ch	82	71	ow	479	321
d	508	423	oy	22	16
dh	506	374	p	324	285
dx	286	195	q	271	223
eh	708	405	r	736	380
el	97	87	s	853	697
em	20	13	sh	119	60
en	90	68	t	943	580
eng	3	2	th	143	127
er	404	335	uh	261	152
ey	342	200	uw	234	184
f	289	199	ux	85	64
g	208	156	v	323	212
h#	1501	835	w	490	296
hh	220	132	y	309	226
ih	1218	421	z	434	248
ix	413	737	zh	29	7

Table 5.1: Count of Phones in the Training and Test Set[2]

[2]Note that there are more instances of the phone [ix] in the test data than in the training data. Since I was unable to find any information regarding the division of the data into test and training sets, I conclude that this irregularity may have been intentional toward a specific goal of the project, or it may have been strictly by chance.

The experimental design aligned observed surface form phones to the canonical form phone. Instances where they did not align perfectly happened for two reasons. While the majority of the cases were due to expected pronunciation variability, there were some cases where there was a temporal misalignment in the source data, particularly at word boundaries. Within the framework of this experiment, it was not possible to definitively distinguish between the two, but I was able to calculate an estimate of phone-level pronunciation variation by first tabulating the number of 'incorrectly' mapped phones, where common allophones were mapped to the base form phone[3] , then tabulating a count of all cases where there was not an exact match between the surface form and the base form phone. The difference in these counts is a rough estimate of the amount of pronunciation variation for each base form phone. There are some problems with this approach, specifically that there are cases of legitimate pronunciation variation that fall outside of the replacement of an allophone for its corresponding canonical phone, particularly among vowels, that are not counted here. These numbers may also downplay pronunciation variation in that the experiment used only the first, and presumably the most common canonical pronunciation, where the CMU Pronouncing Dictionary returned multiple variant pronunciations, so many common and legitimate cases of pronunciation variation were not factored into the experiment.

Table 5.2 shows the base form phones with the highest instance of pronunciation variability. Not surprisingly, the majority of these phones are vowels, which show a greater degree of pronunciation variation than consonants overall. Among the consonants, prominent on the list are those with a syllabic variant, and those with a reduced or flapped form.

---

[3]Common allophones in this data set include syllabic consonants, [el], [em], [en], [eng], and [er] for [L], [M], [N], [NG], and [R] respectively; [d], [q], and [dx] for [T]; and [dx] and [q] for [D]. In addition to the syllabic variant, there is a glided [L], [lg], and a flapped [N], [nx]. Among the vowels, there is a specific reduced form of [IH], [ix], and a reduced [UW], [ux], as well as the more general form, [ax]. [Greenberg et al., 1996]

Phone	Variability (%)
IH	30.569
AH	24.217
L	23.745
UW	20.854
T	20.439
R	20.423
NG	14.917
N	10.231
D	7.656
AA	5.882
ER	5.714
AY2	5.190
AE	4.015
AO	2.941
OW	2.885
EH	2.548
M	1.772

Table 5.2: Base form Phones with the Highest Estimatated Variability

5.2 Observed Phone Prediction Accuracy

Observed phone prediction accuracy was tested on a set of held-out data. Table 5.3 shows the aggregate accuracy for each Feature Set.

Feature Set	
Hayes	0.770
Livescu	0.757
Stevens	0.765

Table 5.3: Aggregate Phone Prediction Accuracy

Based on accuracy alone, there does not seem to be a lot separating the feature sets from each other. The difference in aggregate accuracy between all three is only 1.4% and the accuracies are identical for 1/3 of the phones. Both the Hayes and the Stevens Feature Sets have the highest, or are tied for highest accuracy for 11

phones, while the Livescu has the highest accuracy for 6 phones. What is noteworthy is that those 6 phones that Livescu Feature Set excels at involve a steady vocal tract configuration.

The phones were ranked by accuracy from highest to lowest. Most phones' ranking was within one or two places for each feature set, and the differences in observed phone prediction accuracy for each phone was, on average 1.7% and no more than 5.3% for any phone. Figures 5.1 and 5.2 show the close correspondence of phone accuracy ranking across the feature sets.

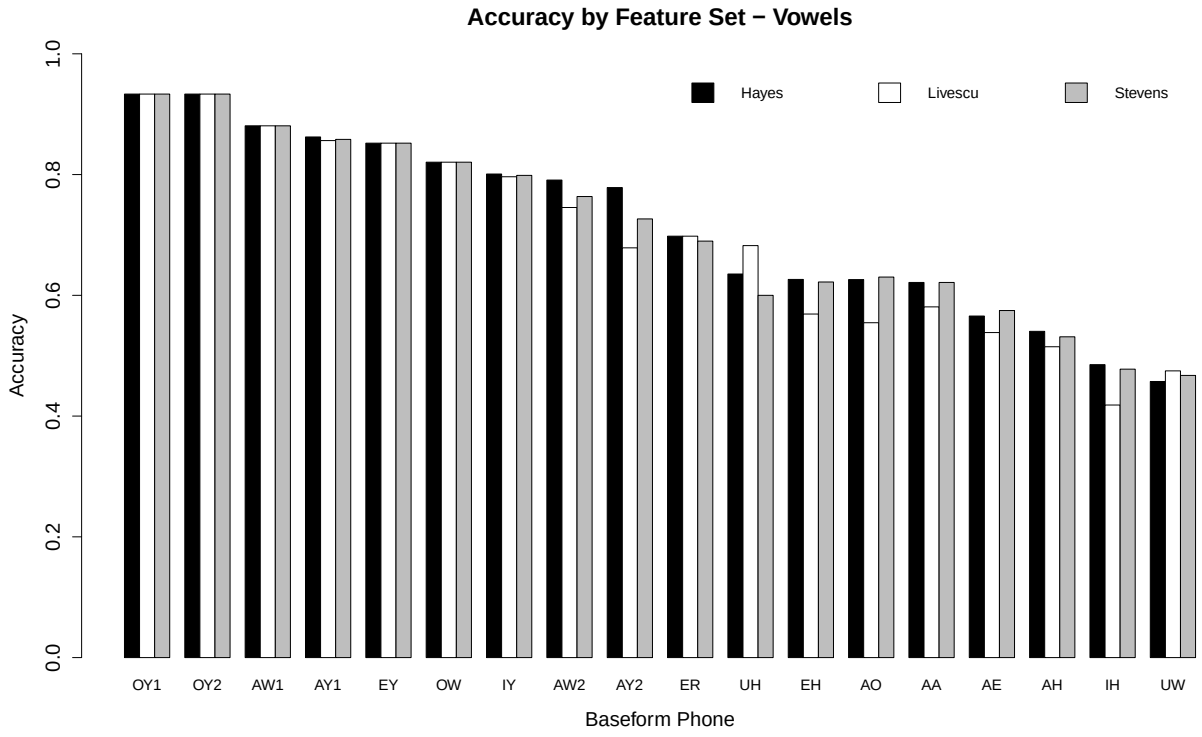

Figure 5.1: Phone Accuracy Ranking by Feature Set - Vowels

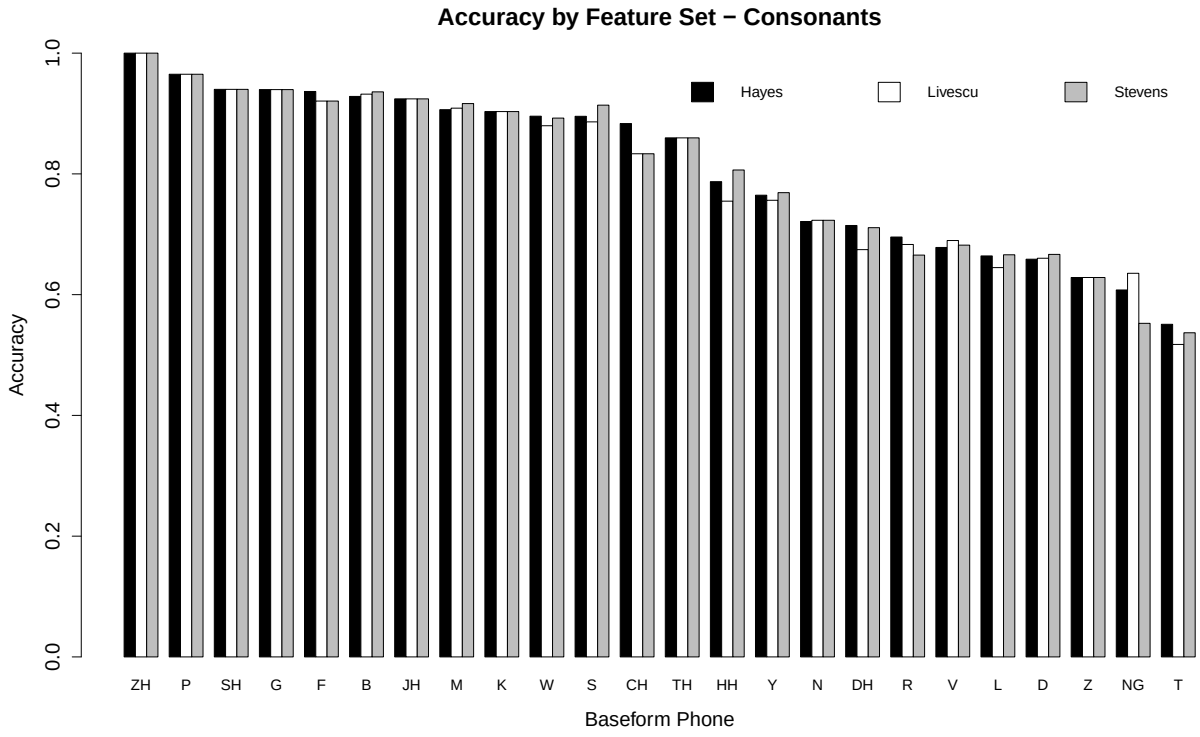

Figure 5.2: Phone Accuracy Ranking by Feature Set - Consonants

There is a strong correspondence between the vowel phones at the lower end of the ranking, and those vowel phones that appear to have the highest degree of phone-level variability, with the exception of [OW], which appears quite robust in spite of its variability. There is not quite as strong a correspondence between phone prediction accuracy ranking and variability among the consonants.

Note that this does not reflect the actual stability of the phones as determined by a perception test or by an ASR system due to some of the characteristics of the data. As seen in table 5.1 some phones do not appear in our data frequently, so the data for those phones should not be considered reliable. Some observed phones in the data were augmented with diacritics, the effects of which were described in the respective feature descriptions in chapter 4. These diacritic effects were reflected in the features used to classify the observed phones, but they did not appear in the inventory of

observed phones.

5.2.1 Casual Speech versus Formal Speech

In addition to overall accuracy among the distinctive feature sets, I also wanted to see which of these models performed best at modeling casual speech versus a more formal speaking style.

To determine whether any of the distinctive feature sets were better able to model the range of pronunciation variation common with casual, spontaneous speech, I tested the model trained on the words considered in isolation against two sets of test data. The first was the observed surface form words from the test set in isolation (i.e. no previous phone for word initial phones, and no following phones for word final phones), which simulated casual speech. The second set of test data was comprised of the canonical pronunciations of the same set of words under the same conditions, which simulated formal speech. A feature set with consistently lower accuracy on the regular test set than the canonical pronunciations would indicate less suitability toward the task of modeling casual speech. Note that there was no accommodation for misalignments in the data for this experiment.

Figures 5.3 through 5.8 show the surface form prediction accuracy for this experiment. The raw data is in tables Tables B.2 through B.4 in appendix B.

It is not at all surprising, given that the model was trained on casual speech data, that, for the majority of phones, the accuracy on the casual speech was either higher or equivalent than on the formal speech. Here, again, the feature sets perform similarly. The Hayes Feature Set has higher or equivalent prediction accuracy for 90.48% of the phones, while the Livescu Feature Set is higher or equivalent for 92.86%, and Stevens is higher or equivalent for 95.23%. Of the phones where the formal was predicted more accurately than the casual, most consistent across all feature sets was [NG], which is not surprising since it is frequently reduced to [n] in casual speech.

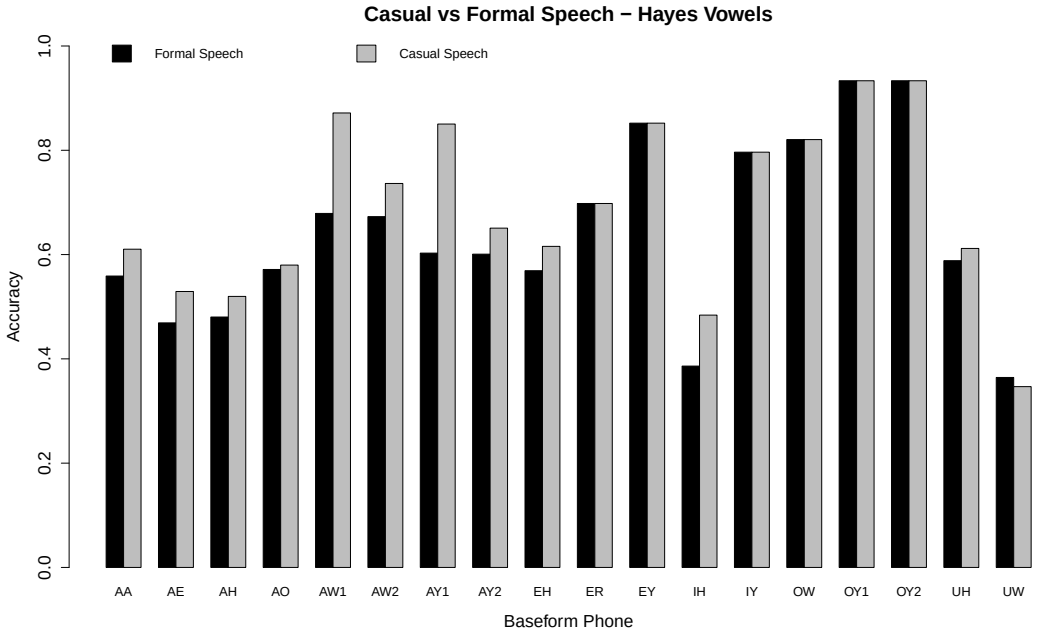

Figure 5.3: Observed Phone Prediction Accuracy for Casual vs Formal Speech - Hayes Vowels

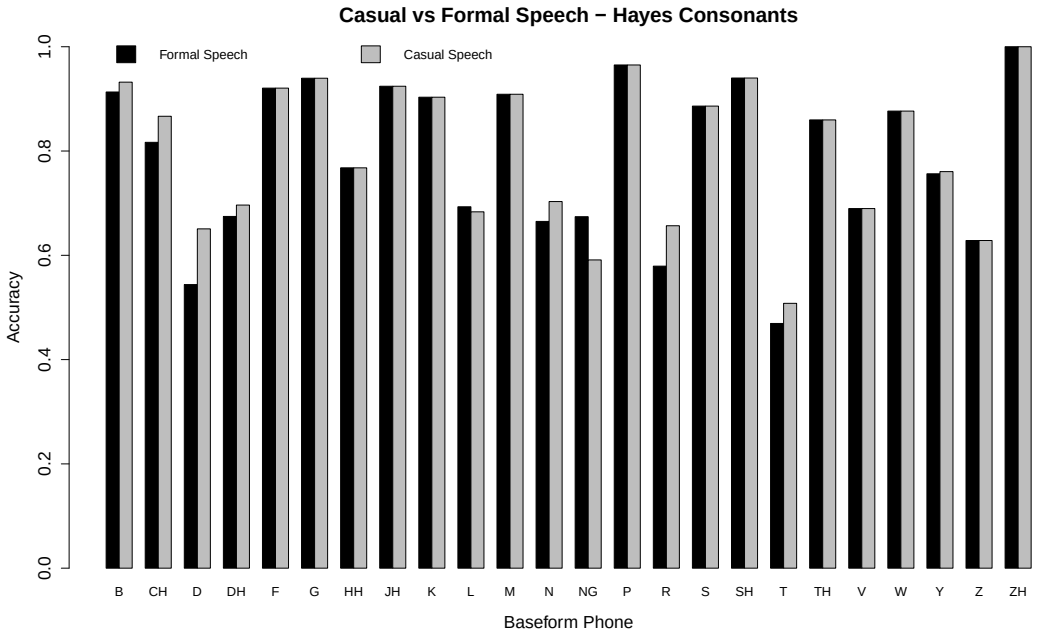

Figure 5.4: Observed Phone Prediction Accuracy for Casual vs Formal Speech - Hayes Consonants

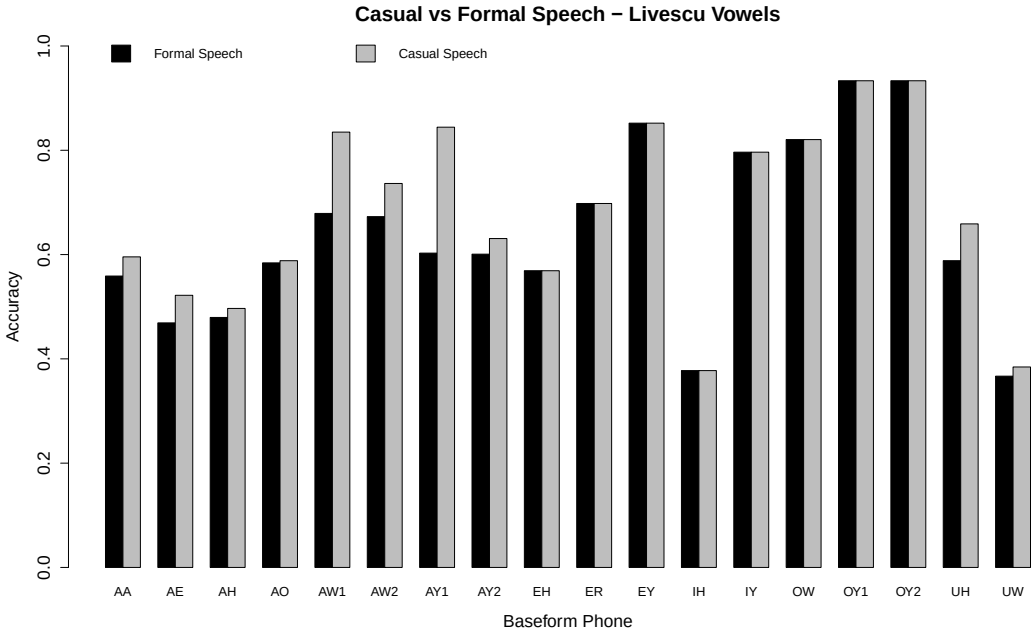

Figure 5.5: Observed Phone Prediction Accuracy for Casual vs Formal Speech - Livescu Vowels

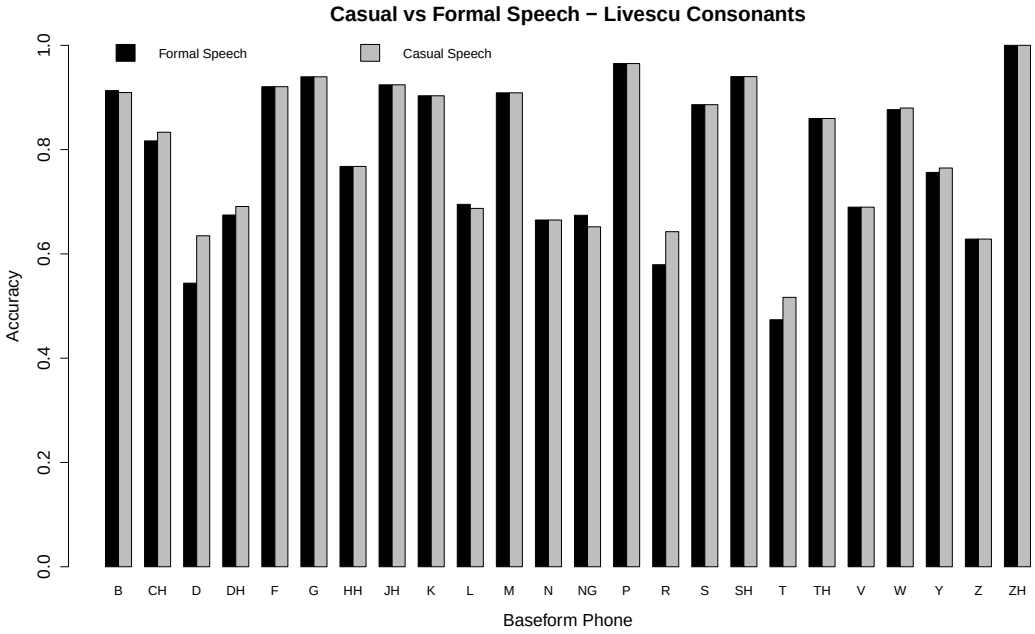

Figure 5.6: Observed Phone Prediction Accuracy for Casual vs Formal Speech - Livescu Consonants

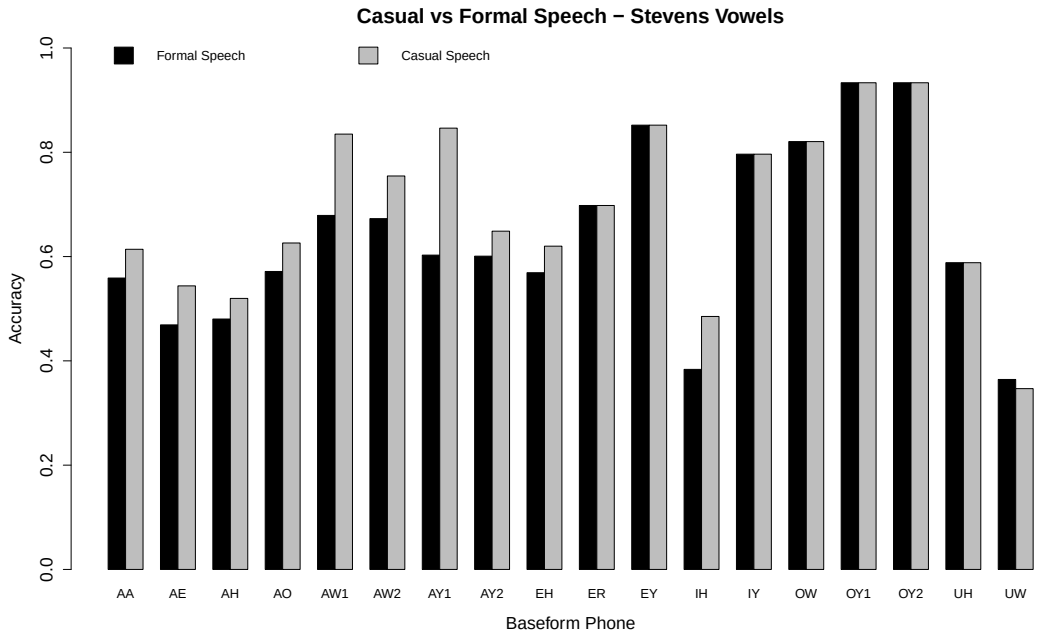

Figure 5.7: Observed Phone Prediction Accuracy for Casual vs Formal Speech - Stevens Vowels

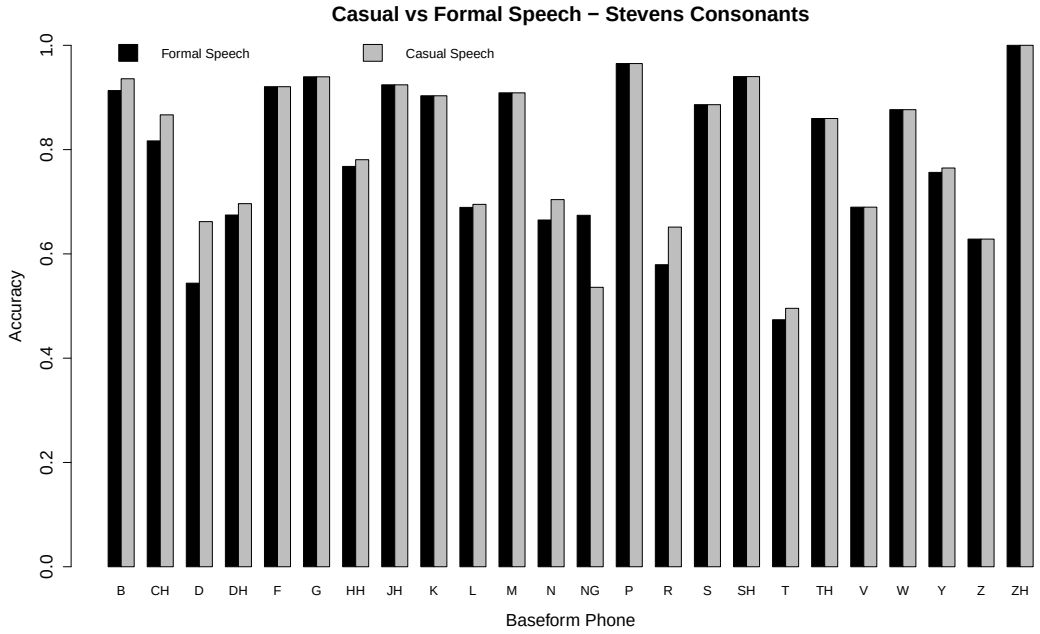

Figure 5.8: Observed Phone Prediction Accuracy for Casual vs Formal Speech - Stevens Consonants

5.3 Decision Tree Characteristics

As described in section 2.4, to classify a test instance using a decision tree requires traversing the tree from the root node to a leaf node, using the test instances's attributes to determine the path. The leaf node at the end of that path represents the instance's predicted classification. In computational terms, this is considered a depth first traversal. For this reason, a shallow tree is preferable to a deeper tree, but tree shallowness is not the sole predictor of efficiency. Another is how much information is required to classify the instance. In the context of a feature-based decision tree, this is represented by non-terminal nodes, which represent decision points. The fewer decision points, the more information is contained within each feature, and the more compact the model. To compare the decision tree models built for each feature set, I am comparing the average depth of the leaf nodes, as well as the number of decision points required to predict the observed surface form. I am also establishing whether there is a relationship between these factors and observed phone prediction accuracy.

5.3.1 Average Depth of Leaf Nodes

Figures 5.9 and 5.10 show the average depth of the leaf nodes for each base form phone decision tree for each feature set. The data is in table B.5 in appendix B.

When you compare these depths to the surface phone prediction accuracy, some patterns become obvious. Primarily, for the trees with a depth of one, the surface phone prediction accuracy matches the accuracy of the mappings of surface forms to canonical forms. These trees with a single node predict a single outcome for every input instance, so from this it is obvious that all surface phones correctly mapped to this canonical base form tree result in an accurate surface form prediction.

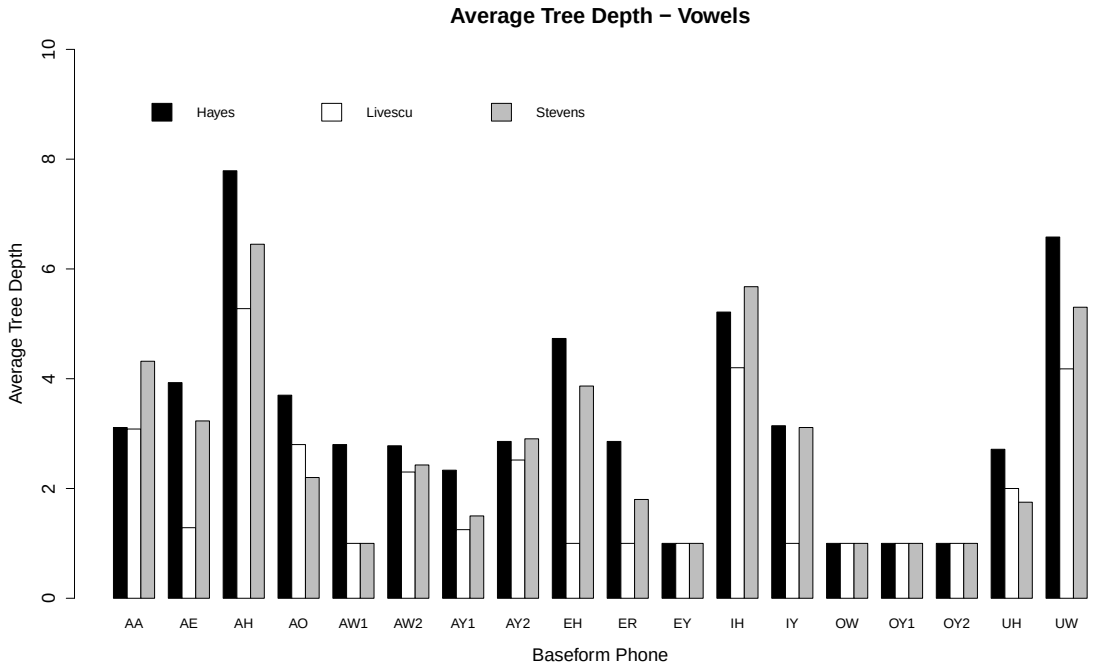

Figure 5.9: Average Depth of Leaf Nodes - Vowels

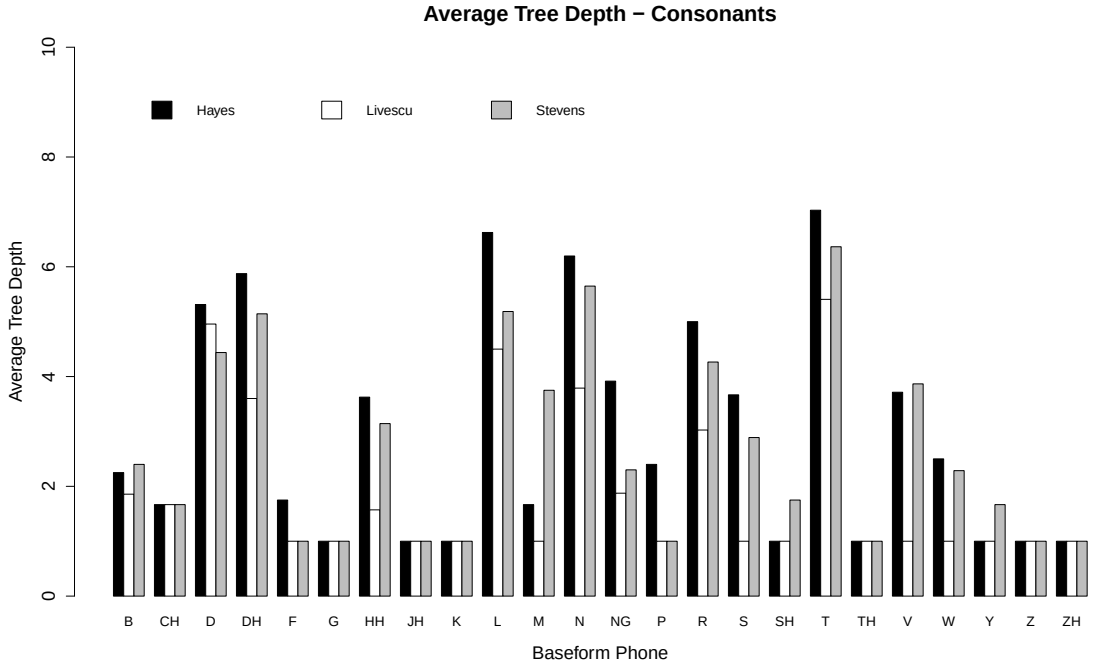

Figure 5.10: Average Depth of Leaf Nodes - Consonants

These figures show that the Hayes Feature Set has the deepest trees for for half of the phones, while the Livescu Feature Set has the shallowest for half. They also reveal a relationship between phone variability and decision tree depth, in that the most variable phones have the deepest decision trees, although, again, this relationship seems to hold stronger for vowels than for consonants.

Due to the difference in the number of features between the Livescu Feature Set, and the Hayes and Stevens Feature Sets, these relative depths do not speak to the superiority of the Livescu Feature Set since it has only eight multi-valued features, which would naturally generate a shallow, wide tree. When you compare the Hayes and Stevens Feature Sets' depths, and recall that the Weka J48 algorithm constructs trees using only the features and provide maximum information gain [Witten et al., 2011], the fact that the average tree depth for the Hayes Feature Set is significantly higher than for the Stevens Feature Set, one can infer that more features are required to predict surface forms and that it is therefore a less compact model.

Figures 5.11, 5.12, and 5.13 show the relationship between decision tree depth and accuracy. The line represents a linear regression model of the data. The consistently negative slopes show a consistently inverse relationship across feature sets: the deeper the tree, the less accurate its predictions. This inverse relationship does not follow from decreasing measures of information gain, however, but is simply an artifact of variability; the more variable a phone, the more difficult it is to predict, requiring deeper trees and resulting in lower accuracy.

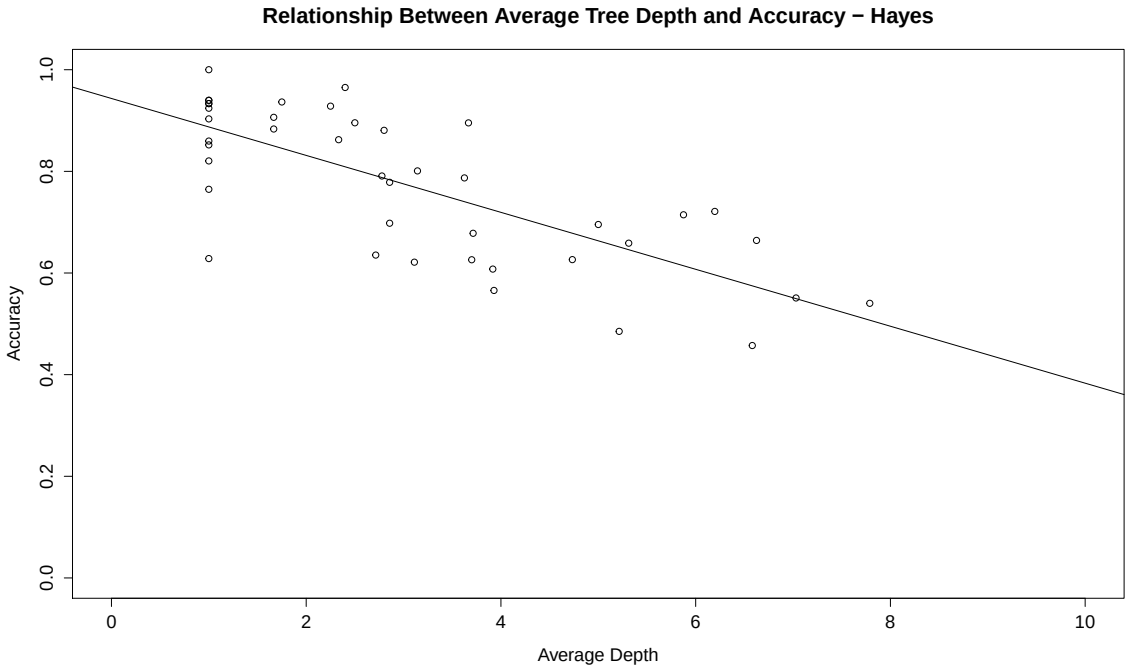

Figure 5.11: Relationship Between the Average Tree Depth and Accuracy - Hayes

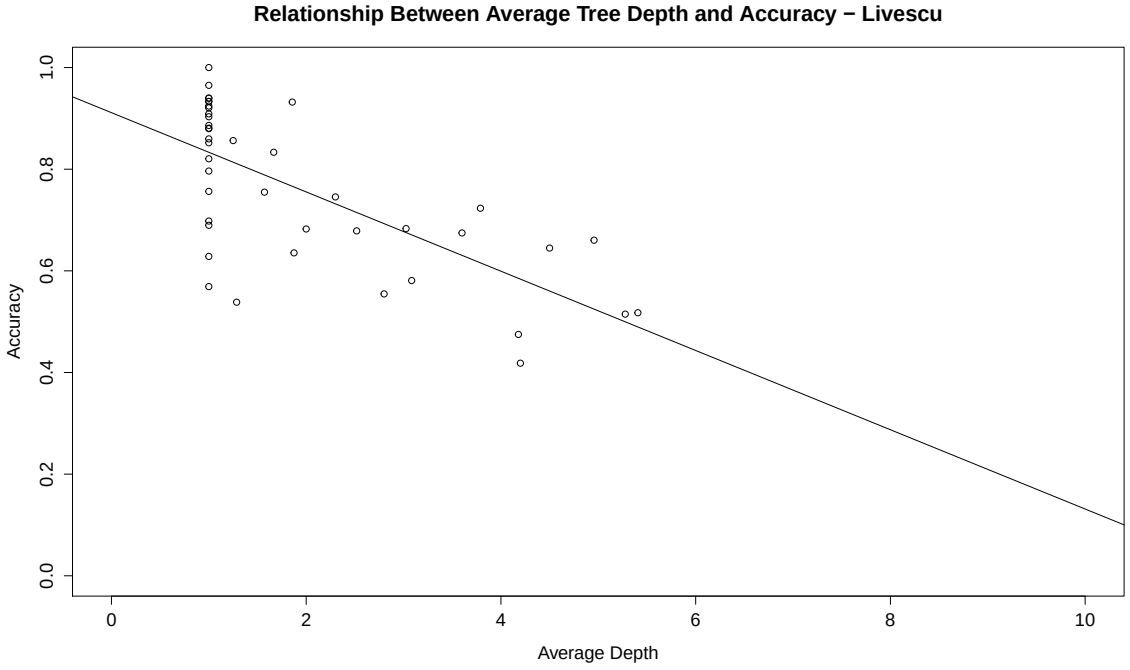

Figure 5.12: Relationship Between the Average Tree Depth and Accuracy - Livescu

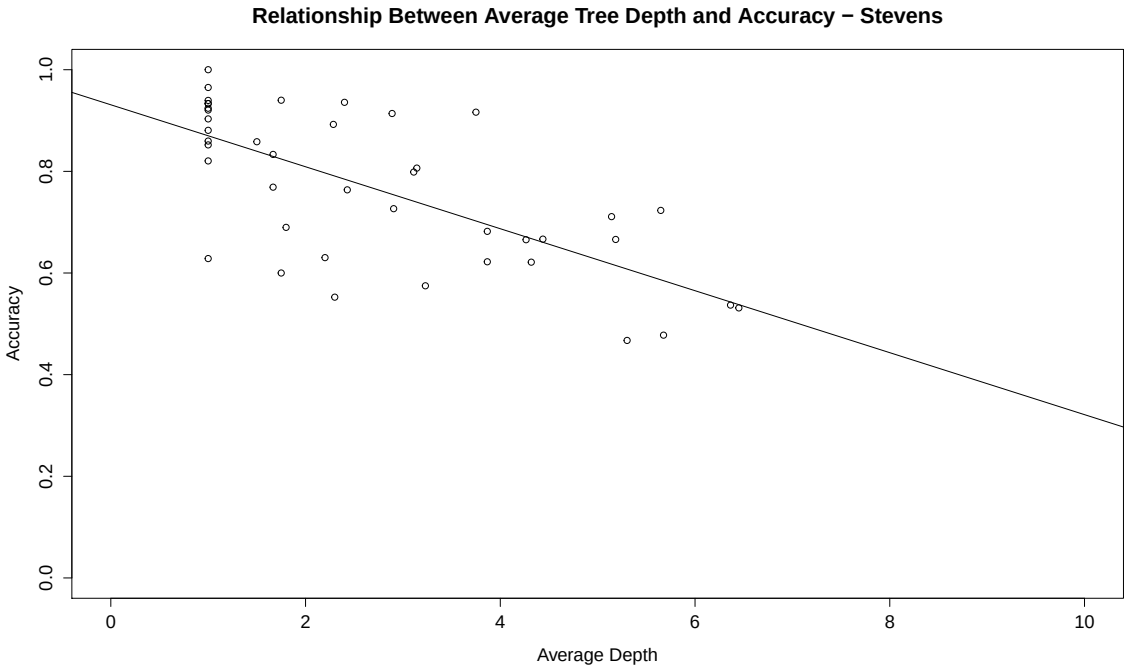

Figure 5.13: Relationship Between the Average Tree Depth and Accuracy - Stevens

5.3.2 Number of Decision Points

In a decision tree, a non terminal node represents a decision point. The fewer decision points, the more compactly the information is stored in a decision tree. Figures 5.14 and 5.15 show the total number of decision points per base form phone for each Feature Set. The data is in table B.6 in appendix B.

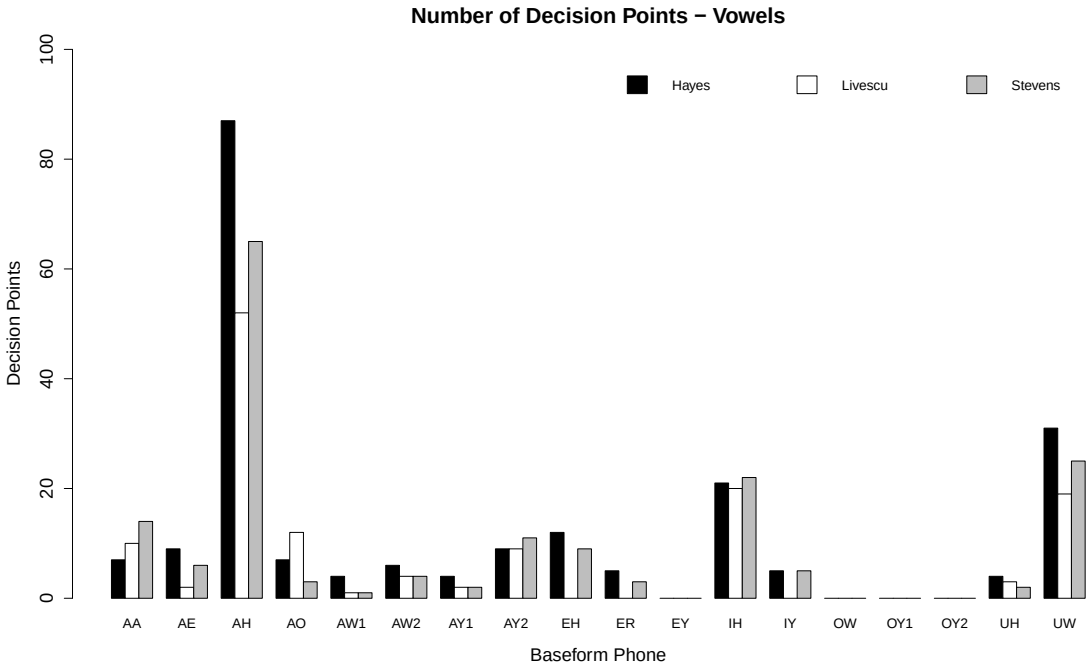

Figure 5.14: Count of Decision Points - Vowels

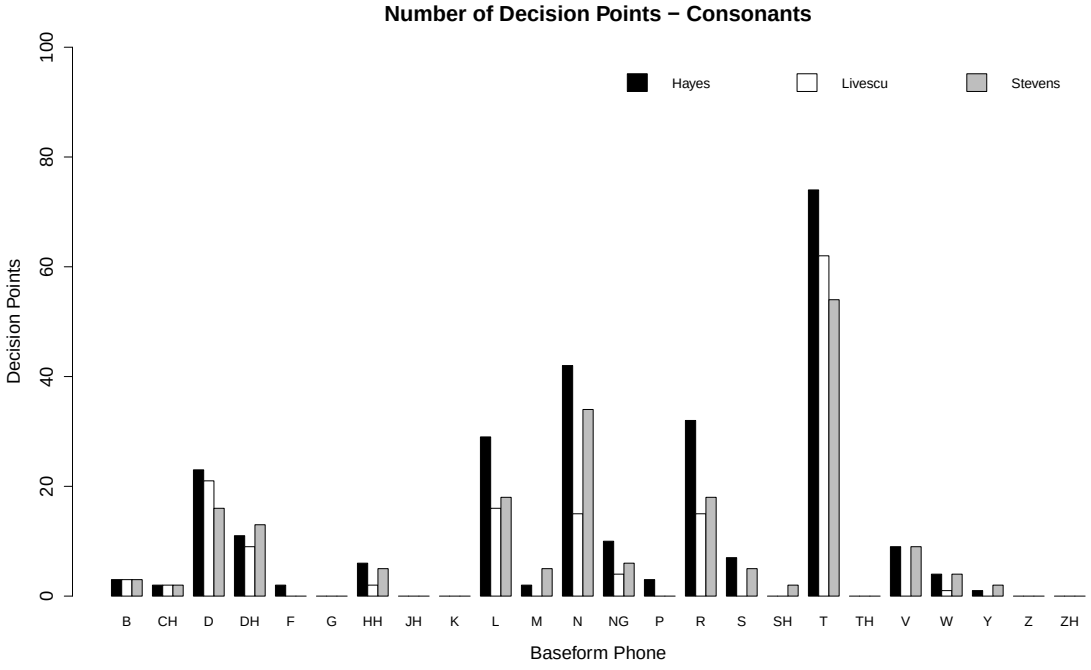

Figure 5.15: Count of Decision Points - Consonants

The Livescu Feature Set has the fewest decision points for 59.5% of the phones. It is able to provide a more compact model due to its multi-valued nature, although it seems unfair to compare it in this regard to a binary feature set. If you compare the two more similar feature sets, the Hayes Feature Set has the highest number of decision points overall for 45.2% of the phones, and is higher than the Stevens Feature Set for 47.6% of phones, sometimes by a considerable amount. Considering that the models only include features that offer significant information gain [Witten et al., 2011], the consistently lower number of decision points indicates that the information is held more compactly for Stevens than for Hayes.

Figures 5.16, 5.17, and 5.18 show the relationship between surface phone prediction accuracy and the number of decision points. The lines represent a linear regression model of the data. The consistently negative slope of the lines speak to the loss of accuracy as the number of decision points increases, not because of the features themselves, but because phones with deeper trees and therefore more decision points are more likely to be highly variable, with multiple allophones and reduced forms, and are more difficult to predict.

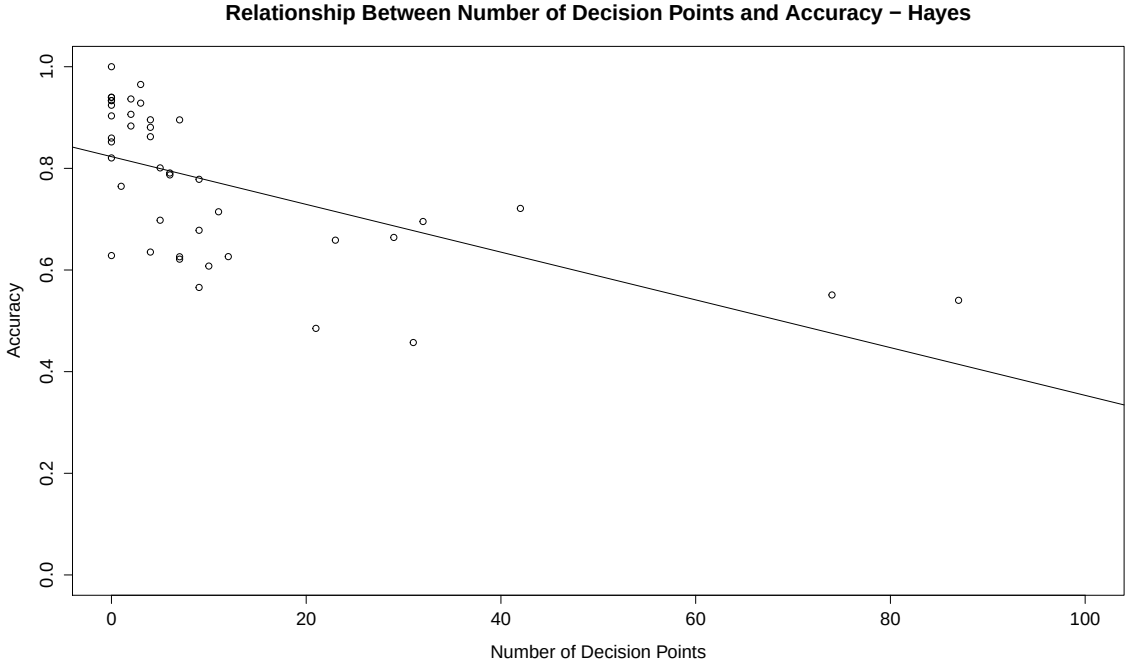

Figure 5.16: Relationship Between the Number of Decision Points and Accuracy - Hayes

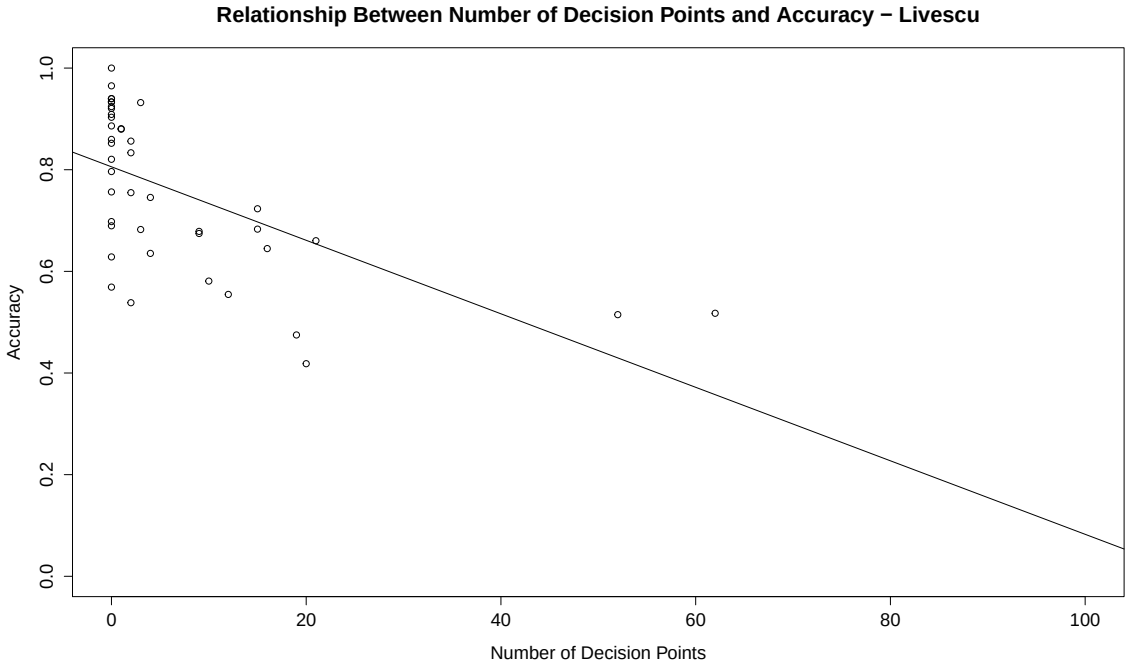

Figure 5.17: Relationship Between the Number of Decision Points and Accuracy - Livescu

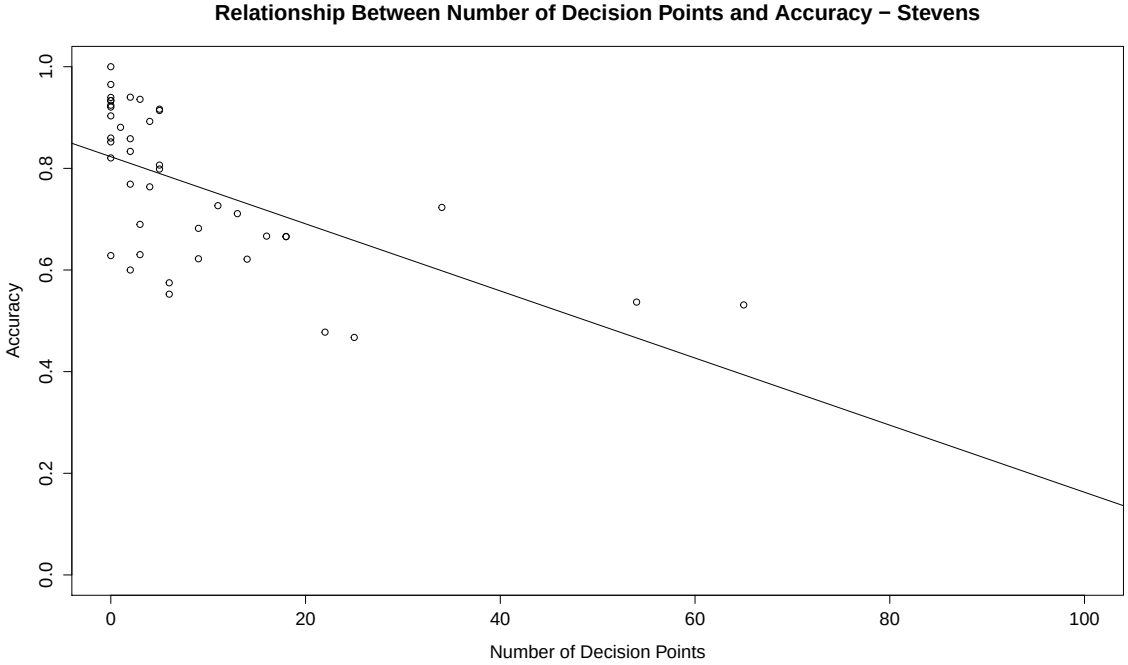

Figure 5.18: Relationship Between the Number of Decision Points and Accuracy - Stevens

5.4 Feature Edit Distance

Feature edit distance is a measure of how close the predicted surface form is to the actual surface form. In terms of distinctive features, I am using a Hamming distance measure that assigns the following values:

- 0 if the features' values are the same
- 0 if one of the feature values is underspecified
- 1 if the feature values are different

I was not able to make it more fine grained due to the differences among the feature sets: the Stevens Feature Set uses a 2-way distinction for unspecified, while the Hayes

Set uses a 1-way distinction, and the Livescu Set has little or no underspecification, so adding a penalty would unduly affect the score of the Stevens Feature Set and provide an unfair advantage to the other two. Not included in these calculations were insertions, deletions, or mappings of a phone to a non-speech sound ('h#'), since insertions and mappings to a non-speech sound are more likely to be the result of an alignment error than an actual observation, and remnants of a deletion often show on an adjacent phone [Ostendorf, 2000] as a diacritic.

Since the different feature sets have a different number of features and a different way of expressing underspecification, I normalized the feature edit distance by the number of features in the set so that it more reflects a measurement of correct feature overlap between the gold standard and the predicted surface form. Table 5.4 shows the average correct feature value overlap for each Feature Set.

Phone	Hayes	Livescu	Stevens	Phone	Hayes	Livescu	Stevens
AA	0.897	0.744	0.872	K	0.701	0.542	0.449
AE	0.795	0.619	0.701	L	0.686	0.705	0.612
AH	0.772	0.708	0.671	M	0.685	0.609	0.601
AO	0.773	0.699	0.810	N	0.732	0.673	0.651
AW1	0.990	1.000	1.000	NG	0.702	0.563	0.545
AW2	0.882	0.661	0.872	OW	0.802	0.589	0.747
AY1	0.797	1.000	1.000	OY1	1.000	1.000	1.000
AY2	0.866	0.743	0.797	OY2	1.000	1.000	1.000
B	0.854	0.700	0.773	P	0.833	0.700	0.539
CH	0.821	0.771	0.667	R	0.723	0.639	0.548
D	0.738	0.740	0.621	S	0.703	0.714	0.607
DH	0.780	0.694	0.636	SH	0.808	0.542	0.636
EH	0.787	0.674	0.700	T	0.665	0.636	0.563
ER	0.714	0.493	0.527	TH	0.808	0.813	0.693
EY	0.904	0.750	0.879	UH	0.833	0.750	0.868
F	0.776	0.600	0.555	UW	0.806	0.588	0.733
G	0.731	0.656	0.500	V	0.645	0.539	0.506
HH	0.785	0.725	0.645	W	0.923	0.688	0.742
IH	0.851	0.661	0.781	Y	0.684	0.672	0.530
IY	0.772	0.700	0.648	Z	0.647	0.695	0.514
JH	0.798	0.844	0.716	ZH	1.000	1.000	1.000

Table 5.4: Average Correct Feature Overlap Between Gold Standard and Predicted Surface Form

The Hayes Feature Set has the highest correct feature overlap for 66.7% of the phones, which was unexpected since it has the most features, and one would think the most opportunity for feature disagreement. This is followed distantly by the Livescu Feature Set, that has the highest correct feature overlap for 19% of phones, and Stevens with the highest for 9.5%. The prediction power of the Hayes Feature Set might lie in the fact that it is a language universal feature set that strives to make distinctions beyond those that are required in English. These non-contrastive features, while not required to differentiate English phones, may provide for a more robust model. This theory is supported by the fact that the Stevens Feature Set, which is more intended to contrast only English phones, has the lowest rate of correct

feature overlap on this data set.

5.5 Phone Class Confusions

Another assessment of these feature sets is how often they confuse phones of different classes. At the highest level, a classifier should be able to distinguish vowels from consonants. Since the C4.5 algorithm uses only the features that are needed [Witten et al., 2011], and due to the greedy nature of the decision tree training process, there will be cases where the gold standard and the predicted surface form differ vastly by phone natural class. Table 5.5 shows the phone class confusions. In this context, the class of sonorants includes vowels, while the sonorant consonants class does not.

Note that where the gold and predicted categories match, a higher percentage is favourable, while where the gold and predicted categories don't match, a lower one percentage shows better performance. Where 'Deletion' appears under Predicted are cases where a surface level input phone is predicted to be a deletion.

There are three interesting interpretations to this experiment. The first is that the Hayes Feature Set is consistently better than the other feature sets at correctly predicting phone class for all phone class pairs. This is possibly due to the non-contrastive features adding a layer of robustness to the model, as I explained in section 5.4. The second interpretation is that the Stevens Feature Set predicts each phone class as a deletion more consistently than the others, although these are at such a low rate that this should not be detrimental to the model. As for rates of phone class confusion, the feature sets performed about equally with the rates differing by less than 1% in most cases, and no more than 1.1%. The overall rate of phone class confusion was less than 6% for all instances.

Gold	Predicted	Hayes	Livescu	Stevens
Consonant	Consonant	0.934	0.931	0.925
Consonant	Vowel	0.027	0.027	0.030
Consonant	Deletion	0.040	0.042	0.045
Vowel	Vowel	0.897	0.888	0.897
Vowel	Consonant	0.060	0.067	0.055
Vowel	Deletion	0.043	0.045	0.049
Sonorant	Sonorant	0.929	0.924	0.928
Sonorant	Obstruent	0.031	0.037	0.027
Sonorant	Deletion	0.039	0.040	0.045
Sonorant Consonant	Sonorant Consonant	0.895	0.890	0.890
Sonorant Consonant	Obstruent	0.038	0.042	0.036
Sonorant Consonant	Deletion	0.033	0.031	0.039
Obstruent	Obstruent	0.917	0.911	0.901
Obstruent	Sonorant	0.039	0.039	0.050
Obstruent	Sonorant Consonant	0.018	0.020	0.024
Obstruent	Deletion	0.044	0.049	0.050
Stop	Stop	0.877	0.856	0.848
Stop	Non Stop	0.023	0.023	0.020
Stop	Deletion	0.059	0.065	0.067
Non Stop	Non Stop	0.928	0.937	0.927
Non Stop	Stop	0.007	0.007	0.008
Non Stop	Deletion	0.030	0.033	0.032

Table 5.5: Phone Class Confusions

Chapter 6

CONCLUSIONS AND FUTURE WORK

In this study, I strove to determine which style of phonetic distinctive feature set was best suited to modeling pronunciation variation, and specifically, the type of pronunciation variation common with casual, spontaneous speech in the context of an Automatic Speech Recognition (ASR) system. Using a non-exhaustive representative set of distinctive feature sets, two articulatory-acoustic hybrids, and one strictly articulatory, I used various measures with which to assess suitability: surface phone prediction accuracy, characteristics of the decision tree models, a measure of feature edit distance, and a measure of phone class confusions.

6.1 Surface Form Prediction Accuracy

Surface form prediction accuracy was assessed under three experimental conditions: words in sequence, treating each speaking turn as a unit and incorporating cross-word influence; words in isolation, treating each word as a separate unit and ignoring cross-word influence; and a simulation of casual versus formal speech. None of the distinctive feature sets was shown to be superior to the others under all conditions: the Hayes Feature Set had the highest aggregate accuracy for the words in sequence experiment, but only by 0.5% over the Stevens Feature Set, and for the words in isolation experiment, the Hayes Feature Set was only 0.01% more accurate than the Stevens and 0.5% more accurate than the Livescu Feature Set. As for the casual versus formal speech experiment, the Stevens Feature Set modeled casual speech best, followed by the Livescu Feature Set, then the Hayes Feature Set. The slight advantage of the Livescu and Stevens Feature Sets for this task may lie in the fact

that their features are based on actual speaker data, Livescu using actual articulatory movements and Stevens using the acoustic signal, while the Hayes Feature Set is based on the vocal tract configuration of a perfect speaker.

6.2 Decision Tree Characteristics

Characteristics of the decision trees generated for each feature set speak to the efficiency of the model as it relates to their compactness. Two such characteristics are decision tree depth and the number of decision points, which are represented here by non-terminal nodes. These factors are even more relevant in this context since the Weka J48 algorithm constructs decision trees using only the features that give maximum information gain [Witten et al., 2011], meaning that the features that appear in the models are those that contain the most classification power.

Here, the Hayes Feature Set tended to have the deepest decision trees. This stems from the fact that the Hayes Feature Set was designed to be language universal and therefore includes features that would make phone distinctions that would not be necessary for the English phone inventory. It consequently defines phones more completely than would be necessary to differentiate between them in our inventory, but this also adds a layer of robustness to the model, as indicated by the higher surface form prediction accuracy for the Hayes Feature Set as compared to the others, as well as the higher rate of correct feature overlap and lower instance of phone class confusions.

The Livescu Feature Set, with considerably fewer features, has the shallowest decision trees, with the aggregate average being a depth of 1.98, compared to 2.72 for the Stevens Feature Set and 3.09 for the Hayes Feature Set, but given the disparity in number of features between it and the other feature sets, I feel this is not a fair comparison. When you consider the average depth and number of features together, it appears that the Stevens Feature Set provides the most compact representation.

The number of decision points is very similar to decision tree depth. Naturally, a deeper tree would have more non-terminal nodes than a shallower tree of similar width. This measure, however, also covers the case where a shallower tree is also wider. Like the decision tree depth measure, the Hayes Feature Set also has the most decision points, sometimes considerably more than the second highest feature set, and the Livescu Feature Set tends to have the least, primarily because it is multivalued. Judging these feature sets for compactness, the Livescu Feature Set would win by its very nature, but in a practical ASR scenario, one wonders if these features could be extracted from the acoustic signal to even begin to make a surface phone prediction, or if it would be more practical to use a less compact model with more, less weighty features, in which case, the Stevens Feature Set would be the more practical compact model.

These measures, however, reveal an issue with using this approach in a practical application. Some of the shallowest, most accurate decision trees have a single decision point, and a single leaf node. Consequently, every input instance generates the same output instance. In this scenario, where I was able to align the phone strings, I was able to direct an observed phone to the most appropriate decision tree model and generate the correct prediction the majority of the time. In a real life scenario, we would not have the benefit of a phone alignment, and every phone sent to these single-depth, single decision point trees would generate a different prediction with equal certainty. We would therefore have to find a means to narrow down the phone candidates before attempting to classify an input instance using these models.

6.3 Feature Edit Distance

Since the feature sets I used differed not only in cardinality, but in how they treated underspecification, a straight measure of feature edit distance would not have been useful. As such, I normalized the feature edit distance by the number of features in each feature set to calculate a measurement of correct feature overlap between

the gold standard phone and the predicted surface phone. Neither of these phone representations incorporate the effects of diacritics, but this affects all feature sets equally.

With this measure, the Hayes Feature Set shows the highest average feature overlap with 79.7%. The Livescu Feature Set follows with 71%, and the Stevens Feature Set has 70.1% correct feature overlap. This is further evidence of the robustness of the Hayes model given to it thanks to the inclusion of features that are potentially redundant and non-contrastive in English.

6.4 Phone Class Confusions

The fact that the Weka J48 algorithm constructs decision trees using only a subset of the features, and that decision tree training is a greedy process, leads to some inevitable cases where the predicted surface form differs from the gold standard input phone in some fundamental way.

Following from the fact that the Hayes Feature Set tended to have highest surface form prediction accuracy for the largest number of phones, it is natural that the Hayes Feature Set would also have the highest rate of correct phone class prediction for most phone class pairs. Where there was confusion between phone classes, it was less than 5% of instances, and all feature sets performed comparably. Interestingly, while all models included a mapping to a deletion as a prediction, the Stevens Feature Set consistently made this prediction incorrectly more often than the other feature sets, although it was less than 5% of all instances for all feature sets.

6.5 Overall Conclusions

Which feature set performed best depends on how much you weigh compactness against robustness. The Livescu Feature Set provided the most compact representation, shown in the depth and number of decision points of the decision tree models, and was not too far behind the leader in terms of surface phone prediction accuracy

(1.3% behind the most accurate), phone class confusions, correct feature overlap, and modeling casual speech. In a practical application, however, I wonder how easily and robustly the required features can be extracted from the acoustic signal as compared to smaller, more granular features which would have more training data and fewer classification labels, and therefore potentially a more robust model [Kirchhoff et al., 2002]. The Hayes Feature Set, on the other hand, was a less compact model, but it showed robustness in every measure of accuracy, correct feature overlap and phone class confusions. The only measure it did not outperform the other feature sets in was the modeling of casual speech in a direct comparison with formal speech, where its aggregate accuracy was 4.5% behind the most accurate feature set. This is likely due to its roots as an articulatory-based feature set where its features are based on the articulatory movement of perfect speech and the inclusion of non-contrastive features affected its accuracy rather than adding robustness to the model.

There are many things I could have done differently that may have greatly affected my results. Encoding any feature dependencies explicitly expressed in the feature set is one of the main ones. As it was, I depended on the machine learning algorithm to learn and incorporate them. Bates (2003) [Bates, 2003] found a marked improvement in accuracy once the feature geometries were modeled, and incorporating them may have differentiated the feature sets better.

6.6 Future Work

As for future work in this direction, I would like to tie this investigation more tightly into the acoustic signal, expand the feature sets studied, and of course, implement the improvements I mentioned above: incorporating feature geometries and some means of narrowing down the candidate phones before sending the features to the decision tree models to be classified. In addition, since we are dealing with the acoustic signal, an acoustic-based phonetic feature set would be better suited to the task than the

articulatory-based ones currently in use. Unfortunately, there do not seem to be any, so work toward that would also be involved.

BIBLIOGRAPHY

[Bartlett et al., 2009a] Bartlett, S., Kondrak, G., and Cherry, C. (2009a). On the syllabification of phonemes. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 308–316. Association for Computational Linguistics.

[Bartlett et al., 2009b] Bartlett, S., Kondrak, G., and Cherry, C. (2009b). Syllabified cmu. `http://webdocs.cs.ualberta.ca/~kondrak/cmudict.html`. [Online; accessed 22-Oct-2013].

[Bates, 2003] Bates, R. A. (2003). *Speaker dynamics as a source of pronunciation variability for continuous speech recognition models*. PhD thesis, University of Washington.

[Benzeghiba et al., 2007] Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., et al. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10):763–786.

[Chin and Dinnsen, 1991] Chin, S. B. and Dinnsen, D. A. (1991). Feature geometry in disordered phonologies. *Clinical Linguistics & Phonetics*, 5(4):329–337.

[Clements and Hallé, 2010] Clements, G. and Hallé, P. A. (2010). "Phonetic bases of distinctive features": Introduction. *Phonetic Bases of Distinctive Features*, 38(1):3–9.

[Cohen et al., 2004] Cohen, M. H., Giangola, J. P., and Balogh, J. (2004). *Voice user interface design*. Addison-Wesley Professional.

[Fosler-Lussier and Morgan, 1999] Fosler-Lussier, E. and Morgan, N. (1999). Effects of speaking rate and word frequency on pronunciations in convertional speech. *Speech Communication*, 29(2):137–158.

[Godfrey et al., 1992] Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

[Greenberg, 1999] Greenberg, S. (1999). Speaking in shorthand–a syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29(2):159–176.

[Greenberg et al., 2003] Greenberg, S., Carvey, H., Hitchcock, L., and Chang, S. (2003). The phonetic patterning of spontaneous american english discourse. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.

[Greenberg et al., 1996] Greenberg, S., Hollenback, J., and Ellis, D. (1996). The switchboard transcription project. In *1996 LVCSR Summer Workshop Technical Reports*.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

[Hayes, 2011] Hayes, B. (2011). *Introductory phonology*, volume 36. Wiley. com.

[Juang and Rabiner, 2005] Juang, B. and Rabiner, L. R. (2005). Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*.

[Jurafsky and Martin, 2000] Jurafsky, D. and Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, volume 2. MIT Press.

[King et al., 2007] King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., and Wester, M. (2007). Speech production knowledge in automatic speech recognition. *The Journal of the Acoustical Society of America*, 121:723.

[Kirchhoff, 2000] Kirchhoff, K. (2000). Integrating articulatory features into acoustic models for speech recognition. *Phonus 5, Institute of Phonetics, University of the Saarland*, pages 73–86.

[Kirchhoff et al., 2002] Kirchhoff, K., Fink, G. A., and Sagerer, G. (2002). Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication*, 37(3):303–319.

[Livescu, 2005] Livescu, K. (2005). *Feature-based pronunciation modeling for automatic speech recognition*. PhD thesis, Massachusetts Institute of Technology.

[Moran, 2012] Moran, S. (2012). Phonetics information base and lexicon. *Seattle: University of Washington dissertation*.

[Ostendorf, 1999] Ostendorf, M. (1999). Moving beyond the 'beads-on-a-string'model of speech. In *Proc. IEEE ASRU Workshop*, pages 79–84.

[Ostendorf, 2000] Ostendorf, M. (2000). Incorporating linguistic theories of pronunciation variation into speech–recognition models. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 358(1769):1325–1338.

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

[Strik and Cucchiarini, 1999] Strik, H. and Cucchiarini, C. (1999). Modeling pronunciation variation for asr: A survey of the literature. *Speech Communication*, 29(2):225–246.

[Witten et al., 2011] Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. Elsevier.

[Wright, 2006] Wright, R. (2006). Intra-speaker variation and units in human speech perception and asr. In *Speech Recognition and Intrinsic Variation Workshop*.

Appendix A

FEATURE CHARTS

A.1 Stevens

This section contains the complete feature matrices as described in section 4.1. Feature values are as follows:

- + means the value is *on* for that feature
- **-** means the value is *off* for that feature
- **x** means the feature is not applicable for that phone
- **0** means the feature is underspecified for that phone

All tables in this section are taken from [Bates, 2003] with modifications.

A.1.1 Vowels

[footnotesize]

ARPABET symbol	iy	ih	ey	eh	ae	aa	ao	ow	ah	uw	ux	uh	ax	ix	er
vocalic	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
consonantal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
continuant	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
sonorant	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
strident	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
delayed release	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
lips	+	+	+	+	+	+	+	+	+	+	+	+	0	0	0
tongue blade	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
tongue body	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
round	-	-	-	-	-	-	+	+	-	+	+	+	0	0	0
anterior	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
distributed	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lateral	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
high	+	+	-	-	-	-	-	-	-	+	+	+	-	0	0
low	-	-	-	-	+	+	+	-	-	-	-	-	-	0	0
back	-	-	-	-	-	+	+	+	+	+	-	+	0	-	0
nasal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
advanced tongue root	+	-	+	-	-	-	-	+	-	+	+	-	-	-	-
constricted tongue root	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-
spread glottis	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
constricted glottis	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
stiff vocal folds	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table A.1: Stevens Distinctive Features for Monopthong Vowels

A.1.2 Glides and Diphthongs

ARPABET symbol	aw		ay		oy		hh	w	y	r	dx	q	nx	hw	lg
vocalic	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-
consonantal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
continuant	+	+	+	+	+	+	+	+	+	+	+	-	+	+	-
sonorant	+	+	+	+	+	+	-	+	+	+	+	-	+	+	+
strident	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
delayed release	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+
lips	+	+	+	+	+	+	0	+	0	0	0	0	0	+	0
tongue blade	0	x	0	x	0	x	0	x	x	+	+	0	+	x	+
tongue body	+	+	+	+	+	+	0	+	x	+	x	0	x	+	+
round	-	+	-	-	+	-	0	+	0	0	0	0	0	+	0
anterior	0	x	0	x	0	x	0	x	x	+	+	0	+	x	+
distributed	0	x	0	x	0	x	0	x	x	-	-	0	-	x	-
lateral	0	x	0	x	0	x	0	x	x	-	-	0	-	x	+
high	-	+	-	+	-	+	0	+	+	x	x	0	x	+	+
low	+	-	+	-	+	-	0	-	-	x	x	0	x	-	-
back	+	+	+	-	+	-	0	+	-	x	x	0	x	+	+
nasal	0	0	0	0	0	0	-	-	-	-	-	-	+	-	-
advanced tongue root	-	-	-	-	-	-	0	+	+	+	0	0	0	+	x
constricted tongue root	+	-	+	-	+	-	0	-	-	-	0	0	0	-	x
spread glottis	-	-	-	-	-	-	+	-	-	-	-	-	-	+	-
constricted glottis	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-
stiff vocal folds	-	-	-	-	-	-	+	-	-	-	-	+	-	+	-

Table A.2: Stevens Distinctive Features for Diphthongs and Glides

A.1.3 Consonants and Corresponding Syllabic Consonants

ARPABET symbol	l	m	n	ng	el	em	en	eng
vocalic	-	-	-	-	+	+	+	+
consonantal	+	+	+	+	+	+	+	+
continuant	-	-	-	-	+	+	+	+
sonorant	+	+	+	+	+	+	+	+
strident	x	x	x	x	x	x	x	x
delayed release	+	-	-	-	+	-	-	-
lips	0	+	0	0	0	+	0	0
tongue blade	+	+	+	x	+	+	+	x
tongue body	x	x	x	+	0	0	0	+
round	0	-	0	0	0	-	0	0
anterior	+	x	+	x	+	x	+	x
distributed	-	x	-	x	-	x	-	x
lateral	+	-	-	x	+	-	-	x
high	x	x	x	+	0	0	0	+
low	x	x	x	-	0	0	0	-
back	x	x	x	+	0	0	0	+
nasal	-	+	+	+	-	+	+	+
advanced tongue root	x	x	x	x	-	-	-	-
constricted tongue root	x	x	x	x	-	-	-	-
spread glottis	-	-	-	-	-	-	-	-
constricted glottis	-	-	-	-	-	-	-	-
stiff vocal folds	-	-	-	-	-	-	-	-

Table A.3: Stevens Distinctive Features for Consonants and Corresponding Syllabic Consonants

A.1.4 Remaining Consonants

ARPABET symbol	v	dh	z	zh	f	th	s	sh	b	d	g	p	t	k	jh	ch
vocalic	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
consonantal	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
continuant	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-
sonorant	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
strident	+	-	+	+	+	-	+	+	x	x	x	x	x	x	+	+
delayed release	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-
lips	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
tongue blade	x	+	+	+	x	+	+	+	x	+	x	x	+	x	+	+
tongue body	x	x	x	x	x	x	x	x	x	x	+	x	x	x	x	x
round	-	-	-	-	-	-	-	-	-	+	-	-	+	-	-	-
anterior	x	+	+	-	x	+	+	-	x	+	x	x	+	x	-	-
distributed	x	+	-	+	x	+	-	+	x	-	x	x	-	x	-	-
lateral	x	-	-	-	x	-	-	-	x	-	x	x	-	x	-	-
high	x	x	x	x	x	x	x	x	x	x	+	x	x	+	x	x
low	x	x	x	x	x	x	x	x	x	x	-	x	x	-	x	x
back	x	x	x	x	x	x	x	x	x	x	+	x	x	+	x	x
nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
advanced tongue root	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
constricted tongue root	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
spread glottis	-	-	-	-	+	+	+	+	-	-	-	+	+	+	-	+
constricted glottis	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
stiff vocal folds	-	-	-	-	+	+	+	+	-	-	-	+	+	+	-	+

Table A.4: Stevens Distinctive Features for Remaining Consonants

A.2 Hayes'

This section contains the complete feature charts as described in section 4.2. Feature values are as follows:

- + means the value is *on* for that feature
- **-** means the value is *off* for that feature
- **0** means the feature is not relevant to the phone in question[Hayes, 2011]

All data for these tables are taken from Hayes (2011) [Hayes, 2011].

A.2.1 Vowels

ARPABET symbol	iy	ih	ey	eh	ae	aa	ao	ow	ah	uw	ux	uh	ax	ix	er
consonantal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
syllabic	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
sonorant	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
continuant	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
delayed release	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
approximant	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
tap	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
trill	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
voice	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
spread glottis	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
constricted glottis	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
labial	-	-	-	-	-	-	+	+	-	+	+	+	-	-	-
round	-	-	-	-	-	-	+	+	-	+	+	+	-	-	-
labiodental	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
coronal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
anterior	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+
distributed	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+
strident	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-
lateral	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dorsal	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
high	+	+	-	-	-	-	-	-	-	+	+	+	-	+	-
low	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-
front	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-
back	-	-	-	-	-	-	+	+	+	+	-	+	-	-	-
tense	+	-	+	-	0	0	-	+	-	+	+	-	-	+	-

Table A.5: Hayes Distinctive Features for Monopthong Vowels

A.2.2 Glides and Diphthongs

ARPABET symbol	aw		ay		oy		hh	w	y	r	dx	q	nx	hw	lg
consonantal	-	-	-	-	+	-	-	-	-	-	+	+	+	-	+
syllabic	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-
sonorant	+	+	+	+	+	+	-	+	+	+	+	-	+	-	-
continuant	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+
delayed release	0	0	0	0	0	0	+	0	0	0	0	-	0	+	+
approximant	+	+	+	+	+	+	-	+	+	+	+	-	+	-	-
tap	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-
trill	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
nasal	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
voice	+	+	+	+	+	+	-	+	+	+	+	-	+	-	-
spread glottis	-	-	-	-	-	-	+	-	-	-	-	-	-	+	-
constricted glottis	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-
labial	-	+	-	-	+	-	-	+	-	-	-	-	-	+	-
round	-	+	-	-	+	-	-	+	-	-	-	-	-	+	-
labiodental	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
coronal	-	-	-	-	-	-	-	-	-	+	+	-	+	-	+
anterior	0	0	0	0	0	0	0	0	0	-	+	0	+	0	+
distributed	0	0	0	0	0	0	0	0	0	+	-	0	-	0	-
strident	0	0	0	0	0	0	0	0	0	-	-	0	-	0	-
lateral	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
dorsal	+	+	+	+	+	+	-	+	+	-	-	-	-	+	-
high	-	+	-	+	-	+	0	+	+	0	0	0	0	+	0
low	+	-	+	-	-	-	0	-	-	0	0	0	0	-	0
front	-	-	-	+	-	+	0	-	+	0	0	0	0	-	0
back	+	+	+	-	+	-	0	+	-	0	0	0	0	+	0
tense	0	+	0	+	-	+	0	+	+	0	0	0	0	+	0

Table A.6: Hayes Distinctive Feature Chart for Diphthongs and Glides

A.2.3 Consonants and Corresponding Syllabic Consonants

ARPABET symbol	l	m	n	ng	el	em	en	eng
consonantal	+	+	+	+	+	+	+	+
syllabic	-	-	-	-	+	+	+	+
sonorant	+	+	+	+	+	+	+	+
continuant	+	-	-	-	+	-	-	-
delayed release	0	0	0	0	0	0	0	0
approximant	+	-	-	-	+	-	-	-
tap	-	-	-	-	-	-	-	-
trill	-	-	-	-	-	-	-	-
nasal	-	+	+	+	-	+	+	+
voice	+	+	+	+	+	+	+	+
spread glottis	-	-	-	-	-	-	-	-
constricted glottis	-	-	-	-	-	-	-	-
labial	-	+	-	-	-	+	-	-
round	-	-	-	-	-	-	-	-
labiodental	-	-	-	-	-	-	-	-
coronal	+	-	+	-	+	-	+	-
anterior	+	0	+	0	+	0	+	0
distributed	-	0	-	0	-	0	-	0
strident	-	0	-	0	-	0	-	0
lateral	+	-	-	-	+	-	-	-
dorsal	-	-	-	+	-	-	-	+
high	0	0	0	+	0	0	0	+
low	0	0	0	-	0	0	0	-
front	0	0	0	0	0	0	0	0
back	0	0	0	0	0	0	0	0
tense	0	0	0	0	0	0	0	0

Table A.7: Hayes' Distinctive Features for Consonants and Corresponding Syllabic Consonants

A.2.4 Remaining Consonants

ARPABET symbol	v	dh	z	zh	f	th	s	sh	b	d	g	p	t	k	jh	ch
consonantal	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
syllabic	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
sonorant	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
continuant	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-
delayed release	+	+	+	+	+	+	+	+	-	-	-	-	-	-	+	+
approximant	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
tap	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
trill	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
voice	+	+	+	+	-	-	-	-	+	+	+	-	-	-	+	-
spread glottis	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
constricted glottis	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
labial	+	-	-	-	+	-	-	-	+	-	-	+	-	-	-	-
round	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
labiodental	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
coronal	-	+	+	+	-	+	+	+	-	+	-	-	+	-	+	+
anterior	0	+	+	-	0	+	+	-	0	+	0	0	+	0	-	-
distributed	0	+	-	+	0	+	-	+	0	-	0	0	-	0	+	+
strident	0	-	+	+	0	-	+	+	0	-	0	0	-	0	+	+
lateral	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dorsal	-	-	-	-	-	-	-	-	-	-	+	-	-	+	-	-
high	0	0	0	0	0	0	0	0	0	0	+	0	0	+	0	0
low	0	0	0	0	0	0	0	0	0	0	-	0	0	-	0	0
front	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
back	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
tense	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A.8: Hayes' Distinctive Features for Remaining Consonants

A.3 Livescu

This section contains the complete feature charts as described in section 4.3. Features and their values are given in table 4.11 in section 4.3

The data for these tables is from Livescu (2005) [Livescu, 2005] with some modifications so that it conforms to the phone set as given in Bates (2003) [Bates, 2003]. In instances in Livescu (2005) where the feature values were presented as a probability distribution, I chose the feature with the highest probability for simplicity.

A.3.1 Vowels

ARPABET	iy	ih	ey	eh	ae	aa	ao	ow	ah	uw	ux	uh	ax	ix	er
LIP-LOC	LAB	LAB	LAB	LAB	LAB	LAB	PRO	PRO	LAB	PRO		PRO	LAB		LAB
LIP-OPEN	WI	WI	WI	WI	WI	WI	WI	WI	WI	NA		WI	WI		WI
TT-LOC	ALV	ALV	ALV	ALV	ALV	ALV	ALV	P-A	ALV	P-A		P-A	ALV		RET
TT-OPEN	M-N	M-N	MID	MID	WI	WI	WI	WI	MID	WI		WI	MID		NA
TB-LOC	PAL	PAL	PAL	PAL	VEL	PHA	PHA	UVU	UVU	VEL		UVU	UVU		UVU
TB-OPEN	NA	M-N	MID	MID	WI	M-N	M-N	M-N	MID	NA		M-N	MID		WI
VEL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL
GLOT	CR	CR	CR	CR	CR	CR	CR	CR	CR	CR	CR	CR	CR	CR	CR

Table A.9: Articulatory Phonology Distinctive Features for Monopthong Vowels

A.3.2 Glides and Diphthongs

ARPABET	aw		ay		oy		hh	w	y	r	dx	q	nx	hw	lg
LIP-LOC	LAB	PRO	LAB	LAB	PRO	LAB	LAB	PRO	LAB	LAB	LAB	LAB	LAB	PRO	LAB
LIP-OPEN	WI	NA	WI	WI	WI	WI	WI	NA	WI	WI	WI	WI	WI	NA	WI
TT-LOC	ALV	P-A	ALV	ALV	ALV	ALV	ALV	P-A	ALV	RET	ALV	0	ALV	P-A	ALV
TT-OPEN	WI	WI	WI	M-N	WI	M-N	MID	WI	M-N	NA	NA	0	NA	WI	CL
TB-LOC	VEL	UVU	PHA	PAL	UVU	PAL	UVU	UVU	PAL	UVU	VEL	0	VEL	UVU	UVU
TB-OPEN	WI	M-N	M-N	M-N	M-N	M-N	MID	NA	NA	WI	MID	0	MID	NA	NA
VEL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL	OP	CL	CL
GLOT	CR	CR	CR	CR	CR	CR	OP	CR	CR	CR	CR	CL	CR	OP	CR

Table A.10: Articulatory Phonology Distinctive Feature Chart for Diphthongs and Glides

A.3.3 Consonants and Corresponding Syllabic Consonants

ARPABET symbol	l	m	n	ng	el	em	en	eng
LIP-LOC	LAB	LAB	LAB	LAB	LAB	LAB	LAB	LAB
LIP-OPEN	WI	CL	WI	W	WI	CL	WI	WI
TT-LOC	ALV	ALV	ALV	P-A	ALV	ALV	ALV	P-A
TT-OPEN	CL	MID	CL	WI	CL	MID	CL	WI
TB-LOC	UVU	UVU	UVU	VEL	UVU	UVU	UVU	VEL
TB-OPEN	NA	MID	MID	CL	NA	MID	MID	CL
VEL	CL	OP	OP	OP	CL	OP	OP	OP
GLOT	CR	CR	CR	CR	CR	CR	CR	CR

Table A.11: Articulatory Phonology Distinctive Features for Consonants and Corresponding Syllabic Consonants

A.3.4 Remaining Consonants

ARPABET	v	dh	z	zh	f	th	s	sh	b	d	g	p	t	k	jh	ch
LIP-LOC	DEN	LAB	LAB	LAB	DEN	LAB	LAB	LAB	LAB	LAB	LAB	LAB	LAB	LAB	LAB	LAB
LIP-OPEN	CR	WI	WI	WI	CR	WI	WI	WI	CR	WI	WI	CR	WI	WI	WI	WI
TT-LOC	ALV	DEN	ALV	P-A	ALV	DEN	ALV	P-A	ALV	ALV	P-A	ALV	ALV	P-A	P-A	P-A
TT-OPEN	MID	CR	CR	CR	MID	CR	CR	CR	MID	CR	WI	MID	CR	WI	CR	CR
TB-LOC	VEL	UVU	UVU	PAL	VEL	UVU	UVU	PAL	UVU	VEL	VEL	UVU	VEL	VEL	PAL	PAL
TB-OPEN	MID	MID	MID	MID	MID	MID	MID	M-N	WI	MID	CR	WI	MID	CR	MID	M-N
VEL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL	CL
GLOT	CR	CR	CR	CR	OP	OP	OP	OP	CR	CR	CR	OP	OP	OP	CR	OP

Table A.12: Articulatory Phonology Distinctive Features for Remaining Consonants

Appendix B

TABLES

B.1 Surface Phone Prediction Accuracy

Table B.1 shows the overall surface phone prediction accuracy.

Phone	Hayes	Livescu	Stevens	Phone	Hayes	Livescu	Stevens
AA	0.621	0.581	0.621	K	0.903	0.903	0.903
AE	0.566	0.538	0.575	L	0.664	0.645	0.666
AH	0.540	0.515	0.531	M	0.906	0.909	0.916
AO	0.626	0.555	0.630	N	0.721	0.723	0.723
AW1	0.881	0.881	0.881	NG	0.608	0.635	0.552
AW2	0.791	0.745	0.764	OW	0.821	0.821	0.821
AY1	0.862	0.856	0.858	OY1	0.933	0.933	0.933
AY2	0.778	0.679	0.727	OY2	0.933	0.933	0.933
B	0.928	0.932	0.936	P	0.965	0.965	0.965
CH	0.883	0.833	0.833	R	0.695	0.683	0.665
D	0.659	0.660	0.667	S	0.895	0.886	0.914
DH	0.715	0.675	0.711	SH	0.940	0.940	0.940
EH	0.626	0.569	0.622	T	0.551	0.518	0.537
ER	0.698	0.698	0.690	TH	0.860	0.860	0.860
EY	0.852	0.852	0.852	UH	0.635	0.682	0.600
F	0.937	0.921	0.921	UW	0.457	0.475	0.467
G	0.940	0.940	0.940	V	0.678	0.690	0.682
HH	0.787	0.755	0.806	W	0.896	0.880	0.892
IH	0.485	0.418	0.478	Y	0.765	0.756	0.769
IY	0.801	0.796	0.799	Z	0.628	0.628	0.628
JH	0.924	0.924	0.924	ZH	1.000	1.000	1.000

Table B.1: Surface Phone Prediction Accuracy

B.1.1 Casual versus Formal Speech

Tables B.2, B.3, and B.4 compare the surface phone prediction accuracy for casual versus formal speech for all feature sets.

Hayes					
Phone	Formal	Casual	Phone	Formal	Casual
AA	0.559	0.610	K	0.903	0.903
AE	0.469	0.529	L	0.693	0.683
AH	0.480	0.520	M	0.909	0.909
AO	0.571	0.580	N	0.665	0.703
AW1	0.679	0.872	NG	0.674	0.591
AW2	0.673	0.736	OW	0.821	0.821
AY1	0.603	0.850	OY1	0.933	0.933
AY2	0.601	0.651	OY2	0.933	0.933
B	0.913	0.932	P	0.965	0.965
CH	0.817	0.867	R	0.579	0.657
D	0.544	0.651	S	0.886	0.886
DH	0.675	0.696	SH	0.940	0.940
EH	0.569	0.616	T	0.469	0.508
ER	0.698	0.698	TH	0.860	0.860
EY	0.852	0.852	UH	0.588	0.612
F	0.921	0.921	UW	0.364	0.347
G	0.940	0.940	V	0.690	0.690
HH	0.768	0.768	W	0.877	0.877
IH	0.386	0.484	Y	0.756	0.761
IY	0.796	0.796	Z	0.628	0.628
JH	0.924	0.924	ZH	1.000	1.000

Table B.2: Hayes Surface Form Prediction Accuracy - Formal versus Casual Speech

Livescu					
Phone	Formal	Casual	Phone	Formal	Casual
AA	0.559	0.596	K	0.903	0.903
AE	0.469	0.522	L	0.695	0.687
AH	0.479	0.497	M	0.909	0.909
AO	0.584	0.588	N	0.665	0.665
AW1	0.679	0.835	NG	0.674	0.652
AW2	0.673	0.736	OW	0.821	0.821
AY1	0.603	0.844	OY1	0.933	0.933
AY2	0.601	0.631	OY2	0.933	0.933
B	0.913	0.909	P	0.965	0.965
CH	0.817	0.833	R	0.579	0.643
D	0.544	0.635	S	0.886	0.886
DH	0.675	0.691	SH	0.940	0.940
EH	0.569	0.569	T	0.474	0.517
ER	0.698	0.698	TH	0.860	0.860
EY	0.852	0.852	UH	0.588	0.659
F	0.921	0.921	UW	0.367	0.384
G	0.940	0.940	V	0.690	0.690
HH	0.768	0.768	W	0.877	0.880
IH	0.377	0.377	Y	0.756	0.765
IY	0.796	0.796	Z	0.628	0.628
JH	0.924	0.924	ZH	1.000	1.000

Table B.3: Livescu Surface Form Prediction Accuracy - Formal versus Casual Speech

Stevens					
Phone	Formal	Casual	Phone	Formal	Casual
AA	0.559	0.614	K	0.903	0.903
AE	0.469	0.544	L	0.689	0.695
AH	0.480	0.520	M	0.909	0.909
AO	0.571	0.626	N	0.665	0.704
AW1	0.679	0.835	NG	0.674	0.536
AW2	0.673	0.755	OW	0.821	0.821
AY1	0.603	0.846	OY1	0.933	0.933
AY2	0.601	0.649	OY2	0.933	0.933
B	0.913	0.936	P	0.965	0.965
CH	0.817	0.867	R	0.579	0.651
D	0.544	0.662	S	0.886	0.886
DH	0.675	0.696	SH	0.940	0.940
EH	0.569	0.620	T	0.474	0.496
ER	0.698	0.698	TH	0.860	0.860
EY	0.852	0.852	UH	0.588	0.588
F	0.921	0.921	UW	0.364	0.347
G	0.940	0.940	V	0.690	0.690
HH	0.768	0.781	W	0.877	0.877
IH	0.384	0.485	Y	0.756	0.765
IY	0.796	0.796	Z	0.628	0.628
JH	0.924	0.924	ZH	1.000	1.000

Table B.4: Stevens Surface Form Prediction Accuracy - Formal versus Casual Speech

B.2 Decision Tree Characteristics

B.2.1 Average Depth of Leaf Nodes

Table B.5 shows the average depth of the leaf nodes for each base form phone decision tree for each feature set.

Phone	Hayes	Livescu	Stevens	Phone	Hayes	Livescu	Stevens
AA	3.111	3.083	4.318	K	1.000	1.000	1.000
AE	3.929	1.286	3.231	L	6.625	4.500	5.185
AH	7.788	5.277	6.450	M	1.667	1.000	3.750
AO	3.700	2.800	2.200	N	6.196	3.789	5.648
AW1	2.800	1.000	1.000	NG	3.917	1.875	2.300
AW2	2.778	2.300	2.429	OW	1.000	1.000	1.000
AY1	2.333	1.250	1.500	OY1	1.000	1.000	1.000
AY2	2.857	2.519	2.905	OY2	1.000	1.000	1.000
B	2.250	1.857	2.400	P	2.400	1.000	1.000
CH	1.667	1.667	1.667	R	5.000	3.026	4.265
D	5.313	4.957	4.438	S	3.667	1.000	2.889
DH	5.875	3.600	5.143	SH	1.000	1.000	1.750
EH	4.733	1.000	3.867	T	7.031	5.406	6.365
ER	2.857	1.000	1.800	TH	1.000	1.000	1.000
EY	1.000	1.000	1.000	UH	2.714	2.000	1.750
F	1.750	1.000	1.000	UW	6.581	4.179	5.302
G	1.000	1.000	1.000	V	3.714	1.000	3.867
HH	3.625	1.571	3.143	W	2.500	1.000	2.286
IH	5.214	4.200	5.676	Y	1.000	1.000	1.667
IY	3.143	1.000	3.111	Z	1.000	1.000	1.000
JH	1.000	1.000	1.000	ZH	1.000	1.000	1.000

Table B.5: Average Depth of Leaf Nodes

B.2.2 Number of Decision Points

Table B.6 shows the number of decision points, represented by non-terminal nodes, in each decision tree.

Phone	Hayes	Livescu	Stevens	Phone	Hayes	Livescu	Stevens
AA	7	10	14	K	0	0	0
AE	9	2	6	L	29	16	18
AH	87	52	65	M	2	0	5
AO	7	12	3	N	42	15	34
AW1	4	1	1	NG	10	4	6
AW2	6	4	4	OW	0	0	0
AY1	4	2	2	OY1	0	0	0
AY2	9	9	11	OY2	0	0	0
B	3	3	3	P	3	0	0
CH	2	2	2	R	32	15	18
D	23	21	16	S	7	0	5
DH	11	9	13	SH	0	0	2
EH	12	0	9	T	74	62	54
ER	5	0	3	TH	0	0	0
EY	0	0	0	UH	4	3	2
F	2	0	0	UW	31	19	25
G	0	0	0	V	9	0	9
HH	6	2	5	W	4	1	4
IH	21	20	22	Y	1	0	2
IY	5	0	5	Z	0	0	0
JH	0	0	0	ZH	0	0	0

Table B.6: Number of Decision Points per Tree